THE UNIVERSITY *of* EDINBURGH

Integrated modelling and Bayesian inference applied to population and disease dynamics in wildlife: M.Bovis in Badgers in Woodchester Park

Leonardus Jacobus Johannes Zijerveld

THE UNIVERSITY *of* EDINBURGH

Doctor of Philosophy
University of Edinburgh
2012

Abstract

Understanding demographic and disease processes in wildlife populations tends to be hampered by incomplete observations which can include significant errors. Models provide useful insights into the potential impacts of key processes and the value of such models greatly improves through integration with available data in a way that includes all sources of stochasticity and error. To date, the impact on disease of spatial and social structures observed in wildlife populations has not been widely addressed in modelling. I model the joint effects of differential fecundity and spatial heterogeneity on demography and disease dynamics, using a stochastic description of births, deaths, social-geographic migration, and disease transmission. A small set of rules governs the rates of births and movements in an environment where individuals compete for improved fecundity. This results in realistic population structures which, depending on the mode of disease transmission can have a profound effect on disease persistence and therefore has an impact on disease control strategies in wildlife populations. I also apply a simple model with births, deaths and disease events to the long-term observations of TB (*Mycobacterium bovis*) in badgers in Woodchester Park. The model is a continuous time, discrete state space Markov chain and is fitted to the data using an implementation of Bayesian parameter inference with an event-based likelihood. This provides a flexible framework to combine data with expert knowledge (in terms of model structure and prior distributions of parameters) and allows us to quantify the model parameters and their uncertainties. Ecological observations tend to be restricted in terms of scope and spatial temporal coverage and estimates are also affected by trapping efficiency and disease test sensitivity. My method accounts for such limitations as well as the stochastic nature of the processes. I extend the likelihood function by including an error term that depends on the difference between observed and inferred state space variables. I also demonstrate that the estimates improve by increasing observation frequency, combining the likelihood of more than one group and including variation of parameter values through the application of hierarchical priors.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Leonardus Jacobus Johannes Zijerveld*)

To all who kept on believing in me.

Acknowledgements

I would like to thank Glenn Marion and Mike Hutchings for giving me the opportunity to do this PhD and continuing their support even though I spent significant amounts of time during my PhD away from Edinburgh. I am grateful for their positive contributions throughout in which both Glenn and Mike made sure that I understood the ecological and epidemiological applications of my work. Their input has been invaluable in providing the direction in which my reaearch developed. This work would not have been possible without Glenn's input into the mathematical and statistical approach which benefitted greatly from many informal meetings with him.

I would also like to thank Ross Davidson whose alghoritm and code were the starting point for Chapter 2. This provided a great way into programming in C++ and on several occassion he sorted out problems that I had unknowingly introduced in the code. I would also like to thank my roommate at BioSS, Jamie Prenctice, for his company and positive attitude to any problem or question I would present to him. The friendly attitude of the people in BioSS and SAC in general provided for a very positive environment to work in.

I would like to thank Øyvind Fiksen and his colleagues in the Theoretical Ecology Group at the University in Bergen for hosting me after my move to Norway.

I am grateful to BioSS and SRUC(formerly SAC) for funding my PhD and to FERA and the team at Woodchester Park for providing the data, in particular Robbie McDonald who gave us access to the data and Neil Walker who was my point of contact on many occassions.

I would also like to thank Tom Little who was my postgraduate contact at the School of Biological Science at Edinburgh University and Colin Gillespie and Jos Houdijk my examiners.

Finally and most importantly I am thankfull to my wife Patience whose editing skills were invaluable when I was writing up and whose love together with that of our daughters Esther and Saskia kept me going despite whatever else was going on.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis builds on a large volume of work concerned with understanding and modelling ecological and epidemiological processes. Modelling population dynamics has its roots in the classic work of Malthus, Verhulst and Lotka-Volterra and the principles of this type of modelling are outlined by Renshaw (1991). The foundations for the compartmental models commonly used in epidemiology were laid by Kermack and McKendrick (1927). The connection between the ecological and epidemiological models was made by Anderson and May (1979) who realised that epidemiological processes are strongly dependent on fluctuations in population size. Many subsequent advances have been made by considering various sources of population heterogeneity including space (Renshaw, 1991; Tilman and Kareiva, 1997), gender and age (Anderson and May, 1984; Anderson and Trewhella, 1985) which can all have a profound effect on disease dynamics. These approaches have led to ever increasing complexity when modelling natural systems such as the epidemiology of tuberculosis (*Mycobacterium bovis*) in badgers (*Meles meles*). We now face a choice between more complex models that reflect a detailed understanding of the heterogeneity observed in the natural world versus simple models that are more straight forward to interpret and feasible to parameterise with available data sets. Some researchers believe that more complicated models are required if the aim is to use them in a predictive mode (e.g. Swinton et al. (1997)), but an important question is whether such predictions are valid in the absence of sufficient data to constrain the parameter values and their uncertainty. This is an interesting question in itself but also has important implications in the context of TB in badgers as was shown by the results from the Randomised Badger Culling Trial (ISG, 2007) which proved that culling badgers had the effect of increasing the level of disease in cattle as well as the badger population (Jenkins et al., 2007). This is referred to as the "perturbation effect" and was not predicted by any of the existing models (e.g. (White and Harris, 1995a; Smith et al., 1997)) in spite of their complexity.

In this thesis we present and discuss two models that are different in the level of complexity included but use the same modelling approach for stochastic systems. The first is a relatively complex model for the spread of disease in a highly structured population. The outcomes of this type of model are generally hard to predict due to the interaction of the processes included, but by exploring the parameter space we can analyse its behaviour and draw conclusions about

the relative role of each parameter on the outcomes. The second is a simple model for population and disease dynamics which we fit to data from long term ecological and epidemiological observations in order to derive estimates for the parameters and their uncertainty. The utility of dynamic process-based models in developing understanding of complex interacting systems has long been recognized and is further illustrated by the first model. However, the second example serves to demonstrate that the types of models used throughout this thesis are also amenable to rigourous statistical analysis and therefore, if sufficient data is available, can also be used as predictive models.

When modelling natural systems such as wildlife populations, predictions are subject to both variability and uncertainty. In this context, variability refers to the properties of such systems themselves that are inherently variable (e.g the stochastic nature of births, deaths and infection events), variations in the environment from one location to another, and observational uncertainty. Variability is thus an inherent property of the system and should be quantified and represented in models (Wilkinson, 2009), for example by using stochastic processes as we do throughout this thesis. Uncertainty represents our (lack of) knowledge about the system, for example the appropriate value of model parameters or indeed model structure(Davidson et al., 2012). In modelling, we therefore aim to minimise uncertainty as much as possible e.g. by using data efficiently, and then quantify what remains. Although some useful progress in the understanding of a system's behaviour can be made by increasing the complexity (e.g. Chapter 2) it is clear that if models are to be used in a predictive manner then all sources of uncertainty should be quantified and propagated through into quantitative predictions.

The modelling approach we adopt throughout this thesis was initially designed for modelling stochastic chemical kinetics (Cox and Miller, 1965; Gillespie, 1977) but has since been applied to many other systems including ecology and epidemiology (Renshaw, 1991). In this approach, processes are treated as discrete state-space continuous time Markov processes with exponential waiting times. Changes in the state space variables as well as the time between such changes depend only the state immediately before the change occurs. The rates in the models presented here are a function of the size of the state space variables (this is referred to as mass action) and we do not explicitly model the interaction between individuals. However, in the application to population and disease dynamics we are able to register and track life histories of individuals and in Chapter 2 we show how this type of individual based information may be used to assist the interpretation of the results from our models.

The modelling approach used in Chapter 2 also forms an integral part of the inference approach (Chapters 3 and 4), because the Markovian transition probabilities between subsequent states can be used directly in the definition of the event based process likelihood (Gibson and Renshaw, 1998). A generic approach for inference on dynamic models is summarised by Marion et al. (2012). The main components in this approach are the likelihood, the prior and the posterior. The likelihood can be split into (1) the process model and (2) the observation model. In many applications (including those in Chapter 3 and 4) the true state space history is not observed and also needs to be inferred. The observation

model is a probability that quantifies all the uncertainty and error introduced in the "observation process". The prior describes all previous knowledge (e.g. biological theory, data from other sources) about the system of interest. It includes knowledge about both process and observation parameters as well as the model structure although the latter is not inferred in this thesis. Bayes' theorem combines the likelihood and prior to define the posterior probability for all parameters as well as any unobserved or hidden variables and we implement standard techniques to generate samples from it (e.g. Metropolis et al. (1953); Hastings (1970); Geman and Geman (1984); Gelfand and Smith (1990); Green (1995)).

It has been suggested by several authors that the strong spatial social organisation often apparent in wildlife populations can play a significant role in the spread of disease (Cowan and Garson, 1985; White et al., 2003; Shirley et al., 2003; Judge et al., 2006; Vicente et al., 2007). Davidson et al. (2008) presented a model which explores the idea that in many cases such heterogeneous structures are the result of competition for resources and opportunities (including improved reproductive capabilities). This type of heterogeneity can lead to non-random contacts between susceptible and infected individuals (Altizer et al., 2003), which can have a significant impact on disease transmission. In Chapter 2 of this thesis we embed the model by Davidson et al. (2008) within a spatial context and also include a number of alternative modes of disease transmission. In the absence of data for specific wildlife species we systematically explore the parameter space of our model to derive a number of general conclusions about the spread of disease within populations affected by the heterogeneous fecundity structures resulting from competition.

Tuberculosis is endemic in both badgers and cattle in the Southwest of England. The area affected by the disease has expanded since the late 1980s to cover most of the west and south-west of England, and South and mid-Wales. It presents many challenges for farmers and the costs associated with this problem are large (DEFRA (2004) estimated a cost of 1 billion Pounds between 2004 and 2011). In this light it is important to understand the epidemiological processes affecting the badger population and therefore, the data from an undisturbed badger population in Woodchester Park (SW England) that has been studied since the mid 1970s (Delahay et al., 2000; Vicente et al., 2007) provides a unique opportunity to apply our integrated modelling and inference approach. There is a large literature concerned with modelling the problem of TB in badgers (Anderson and Trewhella, 1985; Bentil and Murray, 1993; White and Harris, 1995a; Smith et al., 1995; Barlow, 1996; Swinton et al., 1997; Shirley et al., 2003). These models tend to be complex and are used in a predictive mode to test strategies for controlling the disease (White and Harris, 1995b; Smith et al., 2001). The parameter estimates for these models are based on reviews of existing observations of badger ecology (Anderson and Trewhella, 1985; White and Harris, 1995a; Smith et al., 1997) or on tuning the model until it "fits" observations. In the latter case it is often unclear what constitutes a good fit and none of the parameterisations include any measure of uncertainty.

In Chapters 3 and 4 we present an alternative modelling approach that includes a small number of population and disease dynamics processes in a simple model and combines it with the inference approach we discussed above. It is

designed with the view to applying it to the data from Woodchester Park but could be used with long term observations of any population affected by disease. We start in Chapter 3 with a model for population dynamics processes only. As suggested by Anderson and May (1979) it is important to model population and disease dynamics simultaneously when the population size is not static. This is particularly true for wildlife like badgers that tend to live in small groups of fluctuating size (e.g. Delahay et al. (2000)) and therefore we proceed in Chapter 4 with a model that also includes disease dynamics processes and the aim to infer all parameters simultaneously. The main advances to the inference approach presented in Chapters 3 and 4 of this thesis include (1) the combination of the process likelihood for multiple simultaneously observed groups, (2) the implementation of hierarchical priors to allow variation of parameter values between such groups and (3) the simultaneous estimation of the process and observation parameters which has to date not been routinely applied in continuous time stochastic models and (4) the implementation of a complex observation model of animal capture and subsequent disease testing.

A brief outline of the thesis is given below.

Chapter 2 presents an investigation into the joint effects of differential fecundity and spatial heterogeneity on demography and disease dynamics, using a stochastic description of births, deaths, social and geographic migration, and disease transmission. A small set of simple rules and parameters govern the rates of births and movement in an environment where individuals compete for improved fecundity. We show that this results in a heterogeneous distribution of individuals and their ages across ranks and by exploring a range of disease transmission scenarios we show that this heterogeneity can have a profound impact on disease persistence.

In Chapter 3 we define a much simpler model for population dynamics and fit it to observations by extending an existing integrated approach for Bayesian Markov chain Monte Carlo (MCMC) parameter inference for continuous time discrete state space models. We (a) incorporate an observation model where the observation parameters are also inferred, (b) combine the likelihood for multiple simultaneously observed populations and (c) develop a hierarchical scheme that allows for variation in parameter values between populations. The approach is tested on simulated data and then applied to the long term observations of a population of European badgers in Woodchester Park (SW England).

In Chapter 4 we present a straight forward extension to the process model by including a disease component (SEI). The extension to the observation model is more complex as it involves both the process of of capturing capturing animals as well as two disease testing processes. In this work we estimate population dynamics and disease parameters simultaneously which is particularly relevant in small populations of fluctuating size. We demonstrate that the approach works on simulated data with high disease prevalence.

The final Chapter contains a general discussion of the findings in this thesis.

Chapter 2

Implications of differential fecundity on demography and disease persistence in wildlife populations

2.1 Abstract

We present an investigation into the joint effects of differential fecundity and spatial heterogeneity on demography and disease dynamics, using a stochastic description of births, deaths, social and geographic migration, and disease transmission.[1] Our model comprises a spatially structured population and each group within this population is divided into a number of ranks with identical fecundity. A small set of simple rules and parameters govern the rates of births and movement in an environment where individuals compete for improved fecundity. This results in a range of realistic population structures as reflected in the heterogeneous distribution of individuals and their ages across ranks. The types of structures emerging from our model and our simulations are consistent with those in natural populations and we demonstrate that such heterogeneous structures and the processes that generate them can have a profound effect on disease persistence. When fecundity is not identical across ranks, older (more mature) individuals tend to be concentrated in high fecundity ranks, which leads to an increase in disease persistence compared to the case where fecundity in all ranks is identical. We show that this effect is significant under a number of different scenarios for disease transmission. Our simulations show that the effects of both spatial and fecundity structure on disease dynamics are significant and need to be considered when examining the impact of disease control strategies in wildlife populations.

[1]This chapter was written with the intention of publication in J. Roy. Soc. Interface with co-authors Ross S. Davidson (SRUC), Glenn Marion (BioSS) and Michael R. Hutchings (SRUC)

2.2 Introduction

Individuals within wildlife populations are engaged in competition for resources and opportunities that ultimately result in differential levels of fecundity across the population. Such *differential fecundity* can arise directly from competition for mates or indirectly from spatial segregation within populations between habitats of varying quality. Competitive interactions lead to non-random contacts between susceptible and infected individuals (Altizer et al., 2003), which can have significant impacts on disease transmission. Examples of where this might be important include the transmission of parasites during grooming in mammals (Altizer et al., 2003) and birds (Kulkarni and Heeb, 2007), sexually transmitted diseases (Altizer et al., 2003; Judge et al., 2006), and transmission through the oral faecal route in shared habitats (Judge et al., 2006).

We present a model that incorporates both spatial and fecundity effects and simulations showing that both may be required if we are to accurately capture persistence and spread of infection. The spatial organisation of many wildlife species into stable groups occupying discrete territories is well documented and often there are also strong indications of hierarchical structure within such groups. Some well documented examples of spatial social hierarchies are mongooses (Creel and Rabenold, 1994) and the African wild dog (McNutt, 1996; de Villiers et al., 2003). In the case of the European rabbit it has been suggested that their strong spatial social organisation (Cowan and Garson, 1985) has major implications for the spread of myxomatosis (White et al., 2003) and paratuberculosis (Judge et al., 2006). It has also been suggested that social structure plays a role in the spread of bovine Tuberculosis among European badgers (Shirley et al., 2003; Vicente et al., 2007).

As far as we are aware Davidson et al. (2008) were the first to systematically model and explore the effects of such competitive processes on disease dynamics in wildlife populations. They consider a strongly interacting (local) population which is subdivided according to levels of fecundity and where secondary disease transmission is largely between individuals who share the same level of fecundity. Their results show that in the presence of differential fecundity (via social and/or spatial segregation) competitive interactions lead to increased prevalence and persistence of disease in wildlife populations. However, disease transmission is unlikely to be so restricted in real populations and here we build on this work by extending the analysis and model of Davidson et al. (2008) to explore the impact of a broad range of disease transmission routes. Here we consider the robustness of such differential fecundity effects to secondary infection occurring across all levels of fecundity in the local population. In addition we analyse the impacts of vertical/pseudo vertical transmission as well as the impact of secondary transmission resulting directly from competitive interactions e.g. challenges for territory or position in the social hierarchy.

Both ecological observation (Huffaker, 1958) and theory (e.g. Renshaw (1991); Tilman and Kareiva (1997); Keeling (2000); Keeling and Rohani (2007)) provide evidence that population heterogeneity, and spatial heterogeneity in particular is an important factor in determining the persistence of populations and disease within them. In stochastic models such persistence increases dramatically when

subpopulations are connected e.g. by transmission or movement of individuals between groups. When disease dynamics between subpopulations are asynchronous disease may become extinct in some locations while persisting in others (Keeling, 2000). However, interaction between locations has to be "limited" for heterogeneity to have an effect (Rohani and Ruxton, 1999). When mixing between locations is unlimited, the entire population effectively behaves as one, and the stabilising effect of asynchronicity disappears (Lloyd and May, 1996). Since the stabilising effect of asynchronicity will also disappear in the other extreme where no mixing occurs between groups, there is an optimal amount of mixing which enhances persistence (Keeling, 2000).

In light of this literature we asses the relative impacts of spatial heterogeneity and the dynamics of competition/differential fecundity on the prevalence and persistence of disease. To do so we embed the non-spatial model of Davidson et al. (2008) which represents competitive interactions in the presence of different levels of fecundity within a spatial context. This extends the notion of competitive interactions by taking into account that for an individual to improve their chances of reproductive success they may have to move to another group rather than simply by improving their position within their current group (Creel and Rabenold, 1994; Woodroffe et al., 1995). In our model, the rate and direction of movement of individuals depends on their position within the hierarchical structure and resulting distribution of individuals among ranks is such that the number of individuals in a rank and their ages are correlated to the position of that rank within the hierarchy. This leads to an increase in disease prevalence and persistence compared to the case where fecundity is identical across ranks. When we include the spatial dimension, the resulting population structure is "nested", and allows for different rates of disease transmission within and between spatial groups as well as within and between hierarchical subgroups. This is similar in spirit to the widely studied *household models* common to epidemiology (Ferguson et al., 1997; Ball, 1999), although in our model local disease and population dynamics are considerably more complex and individuals can move between subpopulations.

2.3 Modeling approach

We present a stochastic description of demography and disease dynamics in a structured population. The population is divided into ranks (subgroups of individuals with equal fecundity) and the model describes social interaction and competition driven by differences in fecundity within a spatial context. It includes disease transmission occurring through mass action within sub-populations, in which individuals are either susceptible (S) or infectious (I). A range of fecundity structures (distribution of individuals among ranks with specific fecundity) emerge as a result of different parameterisations of a limited number of simple rules governing births and movements within a spatial-social context.

The model is defined as a continuous time discrete state-space Markov process (Cox and Miller, 1965; Gillespie, 1977) . In this approach we simply define the rates at which events of all types within the model occur. These rates depend on the model parameters as well as the current state of the model, i.e. size of the

population (N) or the number of individuals within each of the possible disease states (S and I). The stochastic process is simulated by drawing the time to the next event from an exponential distribution with parameter given by the sum of the rates of all possible events, and then randomly choosing each event with a probability given by its rate divided by this sum of rates. The model is described in details below and a summary of all event rates and their associated state-space changes and lists of all the parameters used in this chapter are given in Tables 2.2,2.3 and 2.4.Additional information about the implementation of the differential fecundity atlgorithm is given in Appendix D.1.

In the subsequent sections we present a number of simulations with the aim of investigating the effects of differential fecundity on demography and disease dynamics in non-spatial as well as spatial contexts. We use a consistent set of parameters throughout (Table 2.1) and run the simulations for an initial period (200 yrs) after which visual inspection shows that the state-space variables (N, S, I) fluctuate around a stochastic equilibrium. We report the long term steady state averages of the relevant variables over a subsequent period (400 yrs) and also averaged over 10 runs. In the simulations that include spatial movements, we use a 3 by 3 lattice, with periodic boundary conditions in which each territory has a carrying capacity of $K = 100$ and dispersal at identical rates to the four nearest neighbours only.

Parameter	Value or range	Remarks
N_l	5	In figure 2.6 $N_l = 2, 5, 10$
N_g	1	In figure 2.6 $N_g = 9$
r_b	1.0	
K	100	
μ	0.1	
ν	$0.0 - 1.50(0.02)$	Range (increment)
p	1	In figure 2.6 $p = 0$
ϕ_p	$0.0 - 0.35(0.005)$	Range (increment)
μ_p	1.0	In figure 2.1 $\mu_p = 0.1, 1.0$
ϕ_d	500	
Γ	0.05	In figure 2.5a $\Gamma = 0.0$
α	0.0	In figure 2.2 $\alpha = 0.0, 0.1$
p_v	0.0	In figure 2.4 $p_v = 0.0, 0.25$
β_p	0.0	In figure 2.5 $\beta_p = 0.0 - 1.0$

Table 2.1 Summary of parameter values used in the simulations presented in this chapter. See the text and Table 2.4 for an explanation of the meaning of each parameter.

2.4 Differential fecundity affects population size and structure

Our starting point is the hierarchical model for population dynamics presented in Davidson et al. (2008). In their model a single group is divided into N_l ranks. Each rank has a population n_i for $i = 1 \dots N_l$. Differences in fecundity affect the rate of births as well as the rate at which individuals move between ranks in attempts to improve their fecundity. The highest ranks offer reproductive advantages and are the most desirable, but also the most difficult to enter.

Births into a particular rank i from parents in all ranks j are modeled with rate

$$r_b f_i \left(1 - \frac{N}{K}\right) \sum_{j=1}^{N_l} a_j n_j, \tag{2.1}$$

where n_j is the number of individuals in a particular rank j. This rate is controlled by a *per capita* birth rate r_b and a logistic factor which depends on the total group population $N = \sum_{i=1}^{N_l} n_i$ and carrying capacity K. The logistic factor ensures that the total size of a group (in the non-spatial model there is only one) does not exceed its carrying capacity, regardless of any movement between ranks. Differential fecundity is governed by the positive constants a_i and f_i which are defined separately for each rank. The set of constants a_i are the so called "fitness parameters" which increase with rank to reflect the differential fecundity observed in many wildlife species (Dewsbury, 1982; Ellis, 1995) including rabbits (von Holst et al., 2002) and African wild dogs (de Villiers et al., 2003). The constants f_i typically decrease with rank, indicating that young adults are more likely to start off in a low rank (Rödel and von Holst, 2009).

The coefficients a_i and f_i are likely to be highly correlated, and a simple approach is to define them as functions of each other. In Davidson et al. (2008) this was done as follows

$$f_i = a_{N_l - i + 1}, \tag{2.2}$$

so that any trend in a_i across ranks is reversed in f_i, leaving only a_i to be defined. To allow for straight forward comparison with non-hierarchical models the requirement is made that the mean values of each of the hierarchical coefficients over all ranks equals 1. Although a number of other definitions of a_i can be considered, we follow Davidson et al. (2008) in adopting the linear form:

$$a_i = \frac{2(1-c)i}{N_l + 1} + c. \tag{2.3}$$

where c can be varied from 0 for a highly structured situation, to 1 where position within the social structure is not associated with any benefits in terms of reproduction.

Deaths occur in all ranks i at the same rate

$$\mu n_i \tag{2.4}$$

controlled by a natural *per capita* death rate μ

Individuals can move between ranks to improve their reproductive fitness. Two rates are required to fully describe these movements. *All movement attempts* from rank i to all ranks j occur at rate

$$\sum_{j=1}^{N_l} \rho_{ij}(n_i), \tag{2.5}$$

where

$$\rho_{ij}(n_i) = \nu f_i n_i \frac{a_j}{N_l} \tag{2.6}$$

is controlled by a *per capita* movement rate ν and the same hierarchical coefficients a_j and f_i that were used in the birth term. This is justified because ranks with higher reproductive fitness will also be more desirable, and individuals in higher ranks will be less inclined to attempt moving. Given our definition of the parameters a_j and f_i, movement attempts will be predominantly upward (promotion), but downward movements may also occur. The factor a_j/N_l can be interpreted as the probability of attempting entry into rank j. This factor also ensures that the total rate at which individuals leave any given rank does not diverge as the number of ranks in a group increases.

In a competitive environment, only a proportion of movements can be successful. To account for this we define an acceptance probability for movements into rank i as

$$P_p(n_i) = e^{-\phi_p a_i n_i}, \tag{2.7}$$

where the promotion constant $\phi_p \geq 0$ together with the coefficient a_i determines the difficulty of entering a new rank i and thus controls the population of that hierarchical rank. Note that when $\phi_p = 0$ the acceptance probability is 1 and all attempts to move will succeed. When $\phi_p > 0$ some movement attempts fail. So the rate of *all successful movements into* rank i is given by

$$\sum_{j=1}^{N_l} \rho_{ji}(n_j) P_p(n_i). \tag{2.8}$$

The effect of varying the movement rate ν and the constant that determines the probability with which such movement attempts are successful ϕ_p on the size and structure of the population is illustrated in Figure 2.1. The results in Figures 2.1a and 2.1b were generated with simulations where all failed movement attempts result in death. The effect of this additional mortality increases with both the rate of movement ν and the constant ϕ_p, resulting in an increasingly smaller population (Figure 2.1a). The largest equilibrium population occurs when the rate of attempted movement ν is high and most movement attempts are

successful (ϕ_p is small), i.e. large numbers of individuals improve their fecundity, which positively affects the population size.

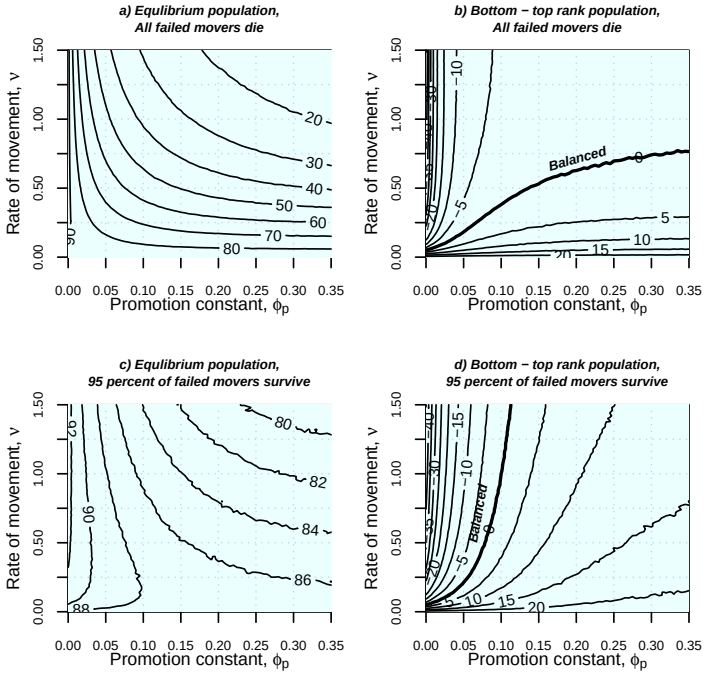


Figure 2.1 Effect of varying rate of movement ν and the promotion constant ϕ_p on the long term equilibrium group population size ($\sum n_i$) and structure. Simulations were run on a group with 5 ranks and a group carrying capacity of $K = 100$, with $r_b = 1.0$, $\mu = 0.1$ and $c = 0$. a,c: Group population in simulations with $c = 0.$, b,d: Difference in population in the lowest and highest rank in a group ($n_1 - n_5$) with $c = 0$. A negative difference indicates that the majority of individuals are in high fecundity ranks, a positive difference indicates that the majority of individuals are in low fecundity ranks. In the top row (a & b) failed movement attempts resulted in death ($\mu_p = 1$), in the bottom row (c & d) 95 % of failed movers survive ($\mu_p = 0.05$).

The distribution of individuals across ranks reflects the distribution of reproductive fitness in the population and is taken as a measure of the structure of a group. Different types of structure arise depending on the movement parameters ν and ϕ_p. We illustrate this in Figure 2.1b, which shows the difference between the equilibrium populations in the lowest and highest ranks at varying conditions. In the case where only the birth process is affected by differential fecundity (i.e no movements, $\nu = 0$), this difference is positive and the number of individuals in high fecundity ranks is small while the majority resides in the lower ranks with lower fecundity. We refer to this as a *bottom-heavy* structure. Increasing the amount of movement between ranks ($\nu > 0$) causes a redistribution of individuals towards the higher ranks with improved fecundity (the corresponding difference in Figure 2.1b decreases and becomes negative). The largest numbers of individuals achieve the best fecundity (*top-heavy*) at high rates of attempted

movement (ν) and low promotion constant (ϕ_p), i.e. most attempts to move succeed. In scenarios where there is no differential fecundity, individuals are evenly distributed across ranks. Note that there are also small differences in the total populations size between models with and without differential fecundity, because when fecundity is not evenly distributed this can lead to overall fecundity to be larger (at high ν, low ϕ_p) or smaller (at low ν).

It was suggested in Davidson et al. (2008) that although mortality subsequent to failed movement attempts in nature can be explained through various mechanisms (Huntingford et al., 1987), excluding the possibility of non-fatal failed movement attempts is a simplification. We extend the model such that only a proportion (μ_p) of failed movement attempts results in death. In this scenario, individuals from rank i that survive a failed movement attempt return to their original rank with rate

$$\sum_{j=1}^{N_l}(1-\mu_p)\left(P_p(n_j)\right)\rho_{ij}(n_i). \tag{2.9}$$

Figures 2.1c and 2.1d show the results of simulations where most individuals survive failed movement attempts ($\mu_p = 0.05$) and return to their original group and rank. In this scenario, the effects on population size and structure are qualitatively very similar to the case where all failed movement attempts result in death. The main difference is that when $\mu_p < 1$, the lower ranks of the model are less depleted, particularly for large ϕ_p when a high proportion of attempts fail (Fig. 2.1d). Reducing mortality for failed movement attempts also has a significant effect on the total population size (Fig. 2.1c), which becomes less suppressed at high ν and ϕ_p.

2.5 Routes of disease transmission in populations with differential fecundity

Having illustrated the effects of differential fecundity on population size and structure we now investigate its implications for disease prevalence and persistence under a number of scenarios for disease transmission in such populations.

Our model applies to disease systems with susceptible (S) and infectious individuals (I) only but could easily be extended to more complex disease dynamics e.g. including latent and removed states. In the most restricted scenario, density dependent infection occurs only between individuals within ranks (subgroups with identical fecundity). This occurs at a *per capita* transmission rate β_1. In a less restricted scenario some transmissions may occur between individuals regardless of where they are within the structure of a group. This group-level, density dependent infection process occurs at a *per capita* transmission rate β_0.

To allow us to compare simulations where the overall strength of infection is identical but the relative importance of each transmission route varies we reparameterise using two new constants: $\Gamma = \beta_0 + \beta_1$ the total strength of transmission and $\alpha = \beta_0/(\beta_0 + \beta_1)$ which controls a continual transition from group-level to

within rank transmission. Given these definitions, *within rank transmission in rank i occurs in addition to group-level transmission at a rate*

$$\Gamma(1-\alpha)S_iI_i \tag{2.10}$$

and group-level infection occurs into rank i at a rate

$$\Gamma\alpha S_i\sum_{j=1}^{N_l} I_j. \tag{2.11}$$

2.5.1 Comparing models with and without differential fecundity

We start our discussion of the effect of various transmission routes with the results of simulations for a population without fecundity benefits ($c = 1$) where infections only occur within the same rank ($\alpha = 0.0$). Figure 2.2a shows how the long term equilibrium prevalence I/N varies over the parameter space that was used in Figure 2.1. At high ν and high ϕ_p disease does not persist within the population (i.e. $I/N = 0$). This is largely due to the fact that in this region of the parameter space the group population is significantly reduced.

Similarity in the shape of the population size contours (Figure 2.1a) and the prevalence contours in Figure 2.2a suggests that there is a close relationship between the group population size N and disease prevalence I/N. This is confirmed by Figure 2.2b but N is not the only factor determining prevalence since, for a given N, the prevalence can take a range of values. When $\alpha = 0$, the only way disease passes between ranks is by the successful movement of infected individuals. Thus higher movement rates result in a better connected disease network and correspond to higher prevalence. Disease is absent when the group size N is small, and below a certain population size, the *disease invasion threshold*, disease can not persist. In this scenario ($c = 1$,$\alpha = 0$) the invasion threshold varies from $N = 57$ at high ν and low ϕ_p to $N = 68$ at low ν and high ϕ_p. These population sizes define the extremes of a range of disease invasion thresholds which are compared with the equivalent ranges for three other scenarios in Figure 2.2d.

The range of ν and ϕ_p over which disease persists increases when differential fecundity is included ($c = 0$, $\alpha = 0$). This increase in persistence occurs regardless of the structure of the group (compare Figures 2.1b and 2.3a). A more subtle difference between these two scenarios is that although for both $c = 1$ and $c = 0$ movements are driven by *population pressure*, the destination in scenarios without differential fecundity ($c = 1$) is random whereas when differential fecundity is included ($c = 0$) individuals in the bottom ranks are more likely to move and they tend to move to the higher ranks. In the latter case the distribution of the ages of individual is always skewed, with a concentration of older (=mature) individuals in the higher ranks (Figure 2.3a,b). Note that the distribution of ages in the entire group is identical whether differential fecundity is included or not (Figure 2.3a). Because older individuals have been exposed to disease for longer they are therefore more likely to have been infected (Figure 2.3c). The concentration of infected individuals in a restricted part of the population results in greater

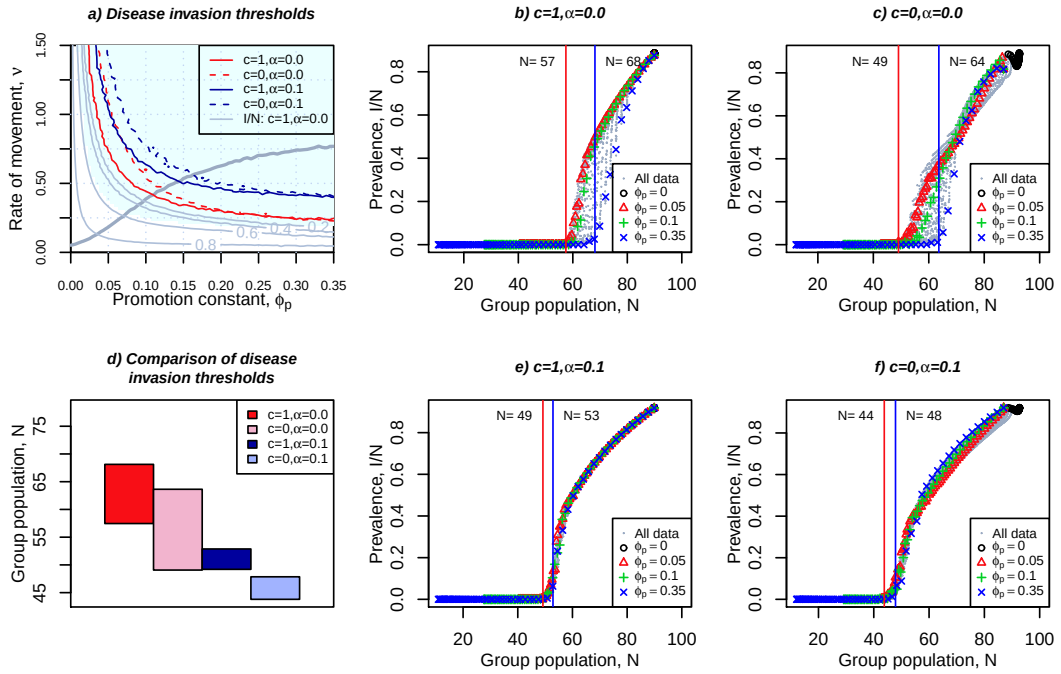


Figure 2.2 Effect of movement between ranks on long term equilibrium disease prevalence and persistence in simulations with transmission within rank only or with some group-level transmission ($\alpha = 0.0$ and 0.1 respectively). These results were generated with total transmission rate $\Gamma = 0.05$, other parameters as in figure 2.1. a: Prevalence (I/N) in a scenario without differential fecundity ($c = 1.0$ and $\alpha = 0.0$) (grey contours). The boundary between areas where $I/N = 0$ and $I/N > 0$ ($I/N = 0.01$) maps combinations of ν and ϕ_p that correspond with the disease invasion threshold. These are included for four scenarios: $c = 1.0$ or 0.0 and $\alpha = 0.0$ or 0.1. The zero contour from 2.1b is also included (grey), indicating parameter combinations at which individuals are evenly distributed across ranks. b,c,e and f: I/N as a function of group population size (N) for the same scenarios as in subfigure (a) and over the entire range of ν and ϕ_p. The width of the data cloud arises because I/N not only depends on N but also on the distribution of individuals across ranks and their ages, which in turn depend on ϕ_p as well as ν. The disease invasion threshold also varies with ν and ϕ_p, and the minimum and maximum value of the invasion threshold in each scenario are marked by red and blue lines. d: Comparison of the ranges of the disease invasion threshold (bounded by min. and max) for the four scenarios considered in this figure.

infectious pressure in such populations as a whole in turn resulting in greater persistence of disease. The effect of the age distribution on the level of disease is strongest in top-heavy populations (high ν, low ϕ_p), where a greater abundance of older individuals in high fecundity ranks coincides with a larger number of individuals residing in those levels (Figure 2.3d). This greater abundance of older individuals in high fecundity ranks is also present in balanced (intermediate

ν and ϕ_p, Figure 2.3e) and also in bottom-heavy populations (low ν and high ϕ_p, Figure 2.3f). but in the latter the abundance of older individuals in higher fecundity ranks is counter balanced by a smaller number of individuals residing in those levels and there is no clear trend of the number of infected individuals with rank. In this case, the total prevalence in scenarios with and without differential fecundity is very similar and the effect on persistence disappears.

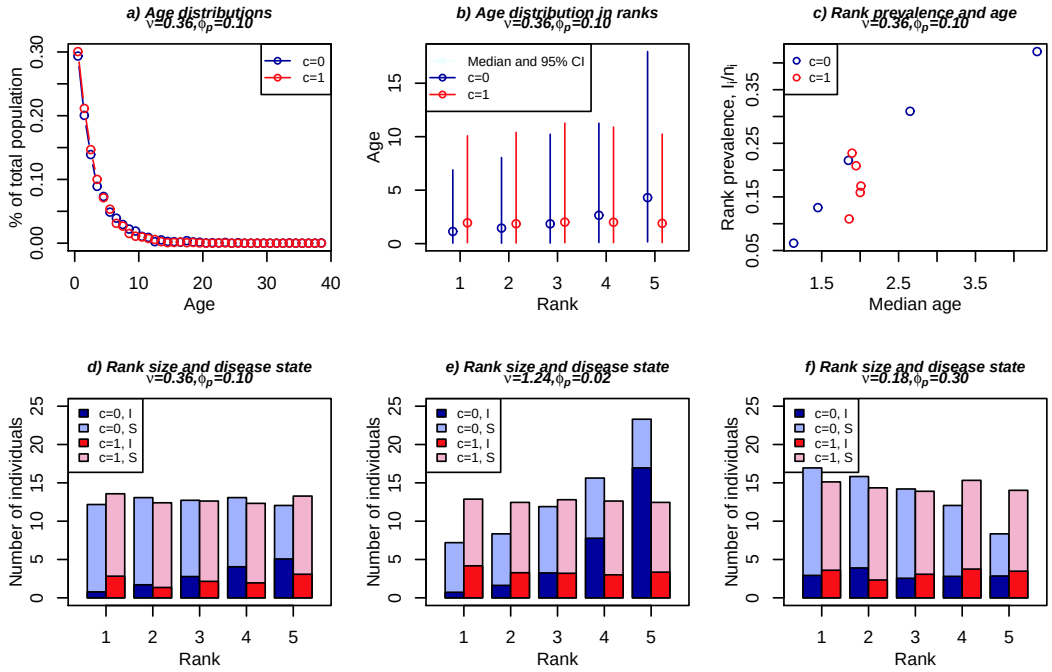


Figure 2.3 Analysis of the relationship between rank, age and disease prevalence in simulations with and without differential fecundity ($c = 0$ and $c = 1$). The data used for this analysis are the dates of births and disease states of those individuals present at the end of a simulation run with specific ν and ϕ_p (as indicated), averaged over 40 runs. All other parameters as in Figure 2.2 with $\alpha = 0$. a: Comparison of the age distributions in the entire population, b: Comparison of median and 95% range of ages across ranks, c: Median age vs. prevalence in ranks (I_i/n_i). The data for a,b and c were generated with $\nu = 0.36$ and $\phi_p = 0.1$. The resulting population structure is balanced when $c = 0$ as can be seen in d, which shows the number of individuals in each rank as well as the number of I and S dark and light shades respectively. Respectively in e and f we show a similar analysis for $\nu = 1.24$, $\phi_p = 0.02$ in which the high fecundity ranks are relatively large and $\nu = 0.18$, $\phi_p = 0.3$ in which the high fecundity ranks are relatively small respectively.

2.5.2 Including group-level transmissions

We now relax the assumption that transmission only occurs between individuals within the same rank ($\alpha = 0$) as this is a very restrictive scenario and in most

realistic situations there will be some interaction between animals in different ranks giving rise to what we refer to above as group-level transmissions ($\alpha > 0$). In the extreme case where all transmissions are at the group-level ($\alpha = 1.0$) the population behaves as one and there are only minor differences between models with and without fecundity benefits, due to small differences in group population size. When $\alpha = 0.1$ there remains a difference between scenarios without ($c = 1$) and with differential fecundity ($c = 0$) suggesting that the effect of differential fecundity remains present even when there are some connections between ranks. In Figures 2.2d-f we show that disease persistence increases by the inclusion of a limited amount of group-level infection ($\alpha = 0.1$). Figures 2.2e-f show a much more direct relationship between N and I/N when $\alpha = 0.1$ than when $\alpha = 0$ (Figure 2.2b-c) and this is confirmed by the narrower ranges of the disease invasion thresholds in Figure 2.2d.

2.5.3 The effect of vertical transmission

We also consider the impact of vertical and pseudo-vertical from infected parent to offspring with probability p_v to reflect the observation that in many wildlife diseases such transmissions provide an important feedback mechanism that plays a key role in disease persistence. The rate at which this occurs is often significant (e.g. in the case of Paratuberculosis in Rabbits, p_v was shown to be 0.33 (Judge et al., 2006), for Toxoplasma Gondii in mice p_v was shown to be as high a 0.75 (Hide et al., 2009)).

An important difference with the group-level transmission route is that pseudo-vertical transmission introduces non-random connections in the transmission network. When differential fecundity is included ($c = 0$) this provides an important feedback between high fecundity ranks that generate most of the young adults and low fecundity ranks where most of the young adults start off. Without differential fecundity ($c = 1$), the additional contacts are essentially random.

Figure 2.4 shows that, when $\Gamma > 0$, pseudo-vertical transmission ($p_v = 0.25$) has a very similar effect to the inclusion of random transmissions between ranks as seen above. Disease persists over a larger range of ν and ϕ_p (Figure 2.4a) and the disease invasion threshold is lower (Figure 2.4b). The ranges of disease invasion thresholds are also narrower (Figure 2.4b), reflecting that the inclusion of p_v introduces new connections between ranks, and causes a more globally connected disease transmission network.

2.5.4 Transmission occurring at movement attempts

In the results so far, movement attempts only directly influence the distribution of disease when an infected individual succeeds in moving. We now consider the additional effect of including a probability of discrete transmission events occurring on all movement attempts. This can be justified, as movement attempts are likely to generate additional contacts between individuals in different ranks and groups. Such contacts are likely to be agonistic and particularly aggressive interactions, and may be very important in the transmission of disease. If we assume that whenever an attempt to move is made, there is one contact with

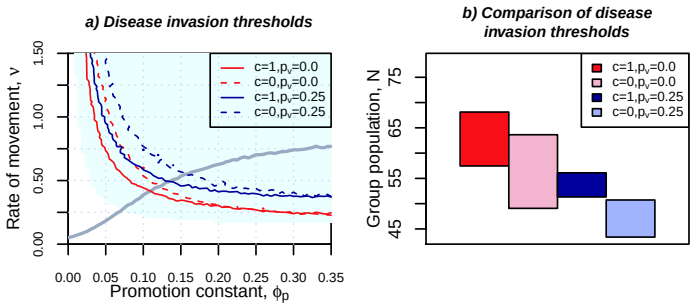


Figure 2.4 Effect of vertical transmission (p_v) on long term equilibrium disease persistence. a: Disease invasion thresholds under four scenarios: with and without differential fecundity ($c = 0.0$ or 1.0), and without and with pseudo vertical transmission ($p_v = 0.0$ or 0.25). Results generated with $\Gamma = 0.05$ and $\alpha = 0.0$ and other parameters as in Figure 2.2. b: Comparison of the ranges of the disease invasion threshold for the four scenarios considered in this figure (see Figure 2.2 and the main text for a further explanation.

an individual in the target group and a specific probability of infection (β_p) we can define the rate of infection of susceptible individuals in rank j caused by attempted moves of infected individuals from rank i to rank j as

$$\nu f_i I_i \frac{a_j}{N_l} \beta_p \frac{S_j}{n_j} \tag{2.12}$$

and the rate of infections of susceptible individuals that succeed in moving from rank i to rank j as

$$\nu f_i S_i \frac{a_j}{N_l} e^{-\phi_p a_j n_j} \frac{I_j}{n_j}. \tag{2.13}$$

Note that we model this as a frequency dependent process where the rates are directly linked to the movement rates. If we include non-fatal failed movement attempts, we need to also consider the rate at which susceptible individuals from rank i survive and attract infection while attempting to move.

In the absence of other modes of disease transmission ($\Gamma = 0$, $p_v = 0$) there is a direct link between disease and the structure that emerges under different values of ν and ϕ_p (compare Figures 2.1b and 2.5a). Trivially, disease is not sustainable at low movement rates. More interestingly, disease does not persist at low values of ν even when all movement attempts involving infectious individuals result in a transmission event ($\beta_p = 1$). When movement rates are sufficient to maintain some disease within the population, simulations including differential fecundity ($c = 0$) have lower persistence than those without ($c = 1$). This can be explained by the observation (from results not shown here) that for most values of ν (except $\nu < 0.3$), the total number of attempted moves when $c = 0$ is less than when $c = 1$. This difference occurs because in the scenario where $c = 0$, at low ν there is a majority in low fecundity ranks and there is a great tendency to move, at

high ν the majority is in high fecundity ranks and the overall tendency to move is low.

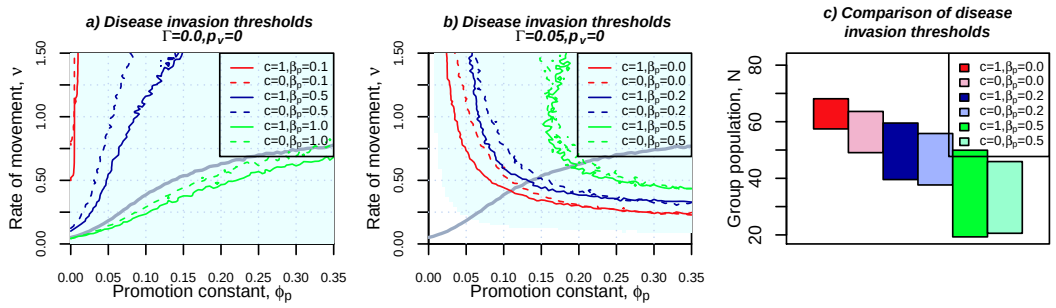


Figure 2.5 Effect of movements between ranks on long term equilibrium disease persistence in simulations with varying probability of transmission associated with movement attempts (β_p). a: Disease invasion thresholds under six scenarios: $c = 1.0$ or 0.0, and $\beta_p = 0.1$ or 0.5 or 1.0. In this comparison there are no other transmissions, i.e. $\Gamma = 0.0$. In this case, disease prevalence is zero in the area below and to the right of the contours. b: Disease invasion thresholds under six scenarios: $c = 1.0$ or 0.0, and $\beta_p = 0.0$ or 0.2 or 0.5, with $\Gamma = 0.05$. Results generated with $\alpha = 0.0$, $p_v = 0.0$ and other parameters as in Figure 2.2. In this case, disease prevalence is zero in the area above and to the right of the contours. c: Comparison of the ranges of the disease invasion threshold for the six scenarios considered in part b of this figure (see Figure 2.2 and the main text for a further explanation).

When both $\Gamma > 0$ and $\beta_p > 0$, the disease invasion contours in Figure 2.5b are a combination of the patterns when $\Gamma > 0$ only (Figure 2.2) and $\beta_p > 0$ only (Figure 2.5a). The main effect of increasing β_p is that the disease becomes more stable (generally persists in smaller populations) and persistence becomes less a function of population size (Figure 2.5c). The inclusion of differential fecundity leads to greater persistence (Figure 2.5b) when β_p is low but when increasing β_p persistence in simulations with high movement rates, the effect can be reversed.

2.5.5 Spatial heterogeneity and differential fecundity

The spatial dimension is implemented as a square lattice consisting of multiple groups with territories of identical carrying capacity. As before, each group is subdivided into a number of ranks, each with specific fecundity. Individuals can move between ranks as well as groups (dispersal). As in the non-spatial model, the motivation to disperse is based on the current position within the spatial-social structure, i.e. individuals in low fecundity ranks and in those ranks with large numbers of individuals have a greater tendency to move. An individual that attempts to move from its current group and rank can move to a new rank in its current group or an alternative group. Local movements, to a different rank within the same group, remain a possibility. The choice of the alternative group depends on the distance from the current group and the choice of rank is

driven by the same parameters as in the non-spatial model (individuals are more likely to attempt moving to higher ranks). The rate of attempted movement in equation (2.6) is changed to a term describing the rate of attempted movement from rank i in group v to rank j in group w:

$$\rho_{vwij}\left(n_{vi}\right) = \nu f_i n_{vi} \frac{a_j}{N_l} F(v,w). \tag{2.14}$$

In this rate, the proportion of all attempted movements from group v to group w is given by the factor

$$F(v,w) = p\delta_{vw} + (1-p)(1-\delta_{vw}) f(d(v,w)), \tag{2.15}$$

where δ_{vw} is the Kronecker delta (i.e. $\delta_{vw} = 1$ when $v = w$ and $\delta_{vw} = 0$ when $v \neq w$), p is the proportion of all movements away from a rank within a group to another rank within the same group (local movement attempt) and $1 - p$ is the proportion of all movements to another group (dispersal)(i.e. when $p = 1$ all attempted movements are local movement attempts, when $p = 0$ all attempted movements are dispersal attempts). The actual distance dependence is described by $f(d(v,w))$, which is some function of the spatial distance between groups v and w. In the simulations presented in this chapter, for computational convenience, we only allow dispersal at equal rates to the four nearest neighbours. Other workers have used a number of alternatives for the distance dependence. Including random dispersal, in which case the rate of dispersal to all other groups is set to be equal, regardless of the distance between them. Another alternative would be to let the rate of dispersal decay as a function of the distance. Commonly used functions are the negative exponential (e.g. Bolker and Pacala (1997)), or a power law function with a negative exponent (e.g. Filipe and Maule (2003)).

To reflect the difference in nature between movements to a new rank within the same group and those that involve movement to a new group we introduce a non-linear acceptance probability for dispersal attempts into group w

$$P_d(N_w) = 1 - \left(\frac{N_w}{K}\right)^{\phi_d}, \tag{2.16}$$

where $\phi_d > 0$ is referred to as the dispersal constant. This function behaves as a logistic function when $\phi_d = 1$ and as a step function as $\phi_d \to \infty$. In the latter case $P_d(N_w) = 0$ when $N_w = K$ and $P_d(N_w) = 1$ when $N_w < K$. This additional acceptance probability also ensures that the carrying capacity of the target group is not exceeded when a movement into it occurs. Combining the acceptance probabilities for social movements and dispersal we get the acceptance probability into rank j of group w for movements from a different group

$$P_a(N_w, n_{wj}) = P_d(N_w) P_p(n_{wj}). \tag{2.17}$$

In the current version of the model there is no disease transmission between individuals in different groups. This results in the least homogeneous mixing of disease and can be considered as an extreme case. In reality, direct spatial disease

transmission as opposed to spread via movement of individuals could play a major role in the spread of disease within a population.

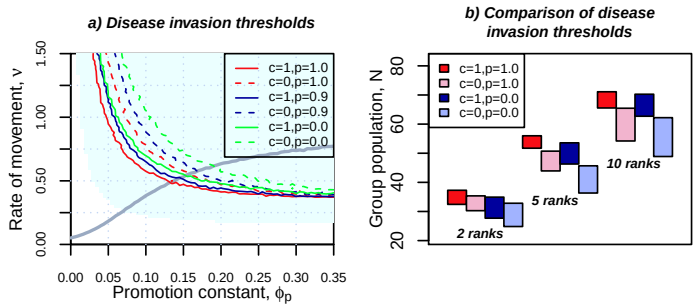


Figure 2.6 Effect of varying rates of movement between groups in a spatial context $(1-p)$ on long term equilibrium disease persistence. a: Disease invasion thresholds under six scenarios: $c = 1.0$ or 0.0, and $p = 1.0$ or 0.9 or 0.0. These results were generated on a 3 by 3 spatial grid, with $\alpha = 0.0$, $p_v = 0.0$, $\beta_p = 0.0$ and all other parameters identical to those used in Figure 2.2. b: Comparison of the ranges of the disease invasion threshold for four of the scenarios considered in part a of this figure (see Figure 2.2 and the main text for an explanation of the meaning of these ranges). The ranges of disease invasion thresholds for the above scenarios but with 2 and 10 ranks are also shown.

Including dispersal stabilizes disease due to the well documented effect of patchiness and asynchronicity in disease dynamics (e.g. Renshaw (1991); Tilman and Kareiva (1997); Keeling (2000)). The spatial effects are of a similar magnitude as the effects of differential fecundity and occur in addition to the latter, as illustrated by Figure 2.6, because simulations with and without fecundity differences ($c = 0$ vs $c = 1$) also have different disease invasion thresholds when spatial movements are included. The relative importance of the effect of spatial movements compared to the effect of differential fecundity depends on the number of ranks within each group (Figure 2.6b). When the number of ranks is small, the inclusion of differential fecundity benefits ($c = 0$) in the absence of spatial movements ($p = 1$) has a smaller effect compared to the inclusion of spatial movements ($p = 0$)in the absence of differential fecundity ($c = 1$). Disease is most persistent in scenarios where spatial movements and differential fecundity are combined. As the number of ranks increases, the effect of differential fecundity becomes increasingly important and outstrips the effect of spatial movements.

2.6 Discussion and conclusions

In this chapter we illustrate the combined effects of differential fecundity and a variety of modes of disease transmission on disease prevalence and persistence. We show that, in most cases, disease persistence increases when differential fecundity drives the demographic processes. This work adds to an extensive literature concerned with the effect of various types of population heterogeneity on disease

dynamics, including space (Renshaw, 1991; Tilman and Kareiva, 1997), sex and age (Anderson and May, 1984; Anderson and Trewhella, 1985) and disease status (Gudelj and White, 2004).

We take a mechanistic approach, focusing on the dynamic nature of the contacts by modeling the processes of movement and dispersal within a structured population driven by fecundity differences. This results in a dynamic contact network and acknowledges that contact networks are at best an approximation to the contact processes occurring within a population (i.e. the processes in which transmission of disease can occur). Building on the non-spatial model of Davidson et al. (2008), we embed the social structure in a spatial context. This takes into account that individuals can improve their chances of reproductive success by moving to a rank with better fecundity within their current group or in a group in another spatial location (Creel and Rabenold, 1994; Woodroffe et al., 1995). The resulting population structure is "nested" This is similar to the *household models* common to epidemiology (Ferguson et al., 1997; Ball, 1999), but differs from them in that in our model individuals can move between subpopulations within the model.

A drawback of simple mass action models (Keeling, 2005) is that they don't include the key features typically observed in human (Read and Keeling, 2003) and animal (e.g. Cross et al. (2004); Lusseau and Newman (2004); Ramos-Fernández et al. (2009)) networks: (1) the number of potential contacts between individuals is limited and highly variable, (2) any two individuals are connected by a small number of steps and (3) individuals are likely to share the same contacts. Although in our model the contact network is not defined explicitly, it does capture some of the key characteristics observed in many networks. By restricting transmission to ranks only we assure that the possible number of contacts between individuals is limited and highly clustered. Allowing individuals to move between ranks and groups may replace the need to explicitly include "small world" properties into the disease transmission network.

Our model should be seen as a conceptual model as to our knowledge there currently is no data that could be used specifically to parameterise the model presented in this chapter. Such data would need to demonstrate a causal link between fecundity differences and population structure in a local and a spatial sense. In addition it would need to demonstrate to what extent the disease transmission network is affected by such structure. In the absence of such data we rely on a systematic investigation of the parameter space in order to assess the potential importance and relative effects of the processes in our model.

Mapping the population structure for varying parameter combinations (Figure 2.1) demonstrates the flexibility of our modeling framework and its ability to represent a range of plausible population structures that reflect the distribution of fecundity across the population.

Differential fecundity and the resulting population structure have a profound impact on disease persistence (Figure 2.3). The key feature in our approach is the dependence of birth and movement rates on fecundity through the parameters a_i, ν and ϕ_p. This generates heterogeneity affecting the number of individuals in each rank as well as the distribution of the ages of individuals across ranks. The concentration of older (mature) individuals in some ranks causes higher preva-

lence in those ranks and greater persistence in the entire group. The simulations presented in this chapter demonstrate that this effect is strongest when disease transmission is restricted to ranks of equal fecundity, but is also present when the contact network is widened by including contacts with other ranks, pseudo-vertical transmission and to some extend transmission on movement attempts.

Including spatial heterogeneity further amplifies the effect of differential fecundity and our simulations suggest that the number of ranks (subgroups with specific fecundity) determines the relative size of fecundity effects compared to that of the purely spatial effects becomes increasingly important as the number of ranks increases. This type of heterogeneity that includes differential fecundity as a driver for spatial movements is different from that in purely spatial models where movements between social groups are purely driven by the population densities in the source and target groups.

In simulations where individuals can survive failed movement attempts and return to their original group and rank, the lower ranks of the model become less depleted and the skew in age distributions is likely to be less pronounced. We have not simulated the effect of differential fecundity on disease dynamics in such scenarios but they are likely to be much more subtle and less pronounced.

Animals living in groups within well defined territories tend to patrol and defend their territorial boundaries (e.g. badgers (Kruuk, 1989)). Such activity can lead to additional contacts (likely to be aggressive in nature) between susceptible and infectious individuals in different social groups, and may be very important in the spread of disease between social groups. Additional contacts may occur when individuals range or forage in neighbouring territories. Disease transmission occurring during these types of contact is currently not included in our model, but in spatial simulations these would lead to a more homogeneous contact structure, an increase in disease prevalence and a less pronounced difference between models with and without differential fecundity.

Given the evidence in nature for the types of hierarchies emerging from our model, the effect of such structures on disease dynamics needs to be considered when examining the impact of control strategies.

Event type	Total event rate	Change in state
Birth into rank i from parent in rank j	$r_b a_j n_j f_i \left(1 - \frac{N}{K}\right)$	$\delta n_i = +1$
Death in rank i	μn_i	$\delta n_i = -1$
Successful move from rank i into rank j	$\nu f_i n_i \frac{a_j}{N_l} e^{-\phi_p a_j n_j}$	$\delta n_i = -1$, $\delta n_j = +1$
Failed movement from rank i into rank j resulting in death	$\nu f_i n_i \frac{a_j}{N_l} \mu_p \left(1 - e^{-\phi_p a_j n_j}\right)$	$\delta n_i = -1$
Failed movement from rank i into rank j not resulting in death	$\nu f_i n_i \frac{a_j}{N_l} (1 - \mu_p) \left(1 - e^{-\phi_p a_j n_j}\right)$	$\delta n_i = 0$

Table 2.2 Event rates and associated changes in state-space for the non spatial model without disease

Event type	Total event rate	Change in state
Birth into rank i from parent in rank j	$r_b a_j n_j f_i \left(1 - \frac{N}{K}\right)$	$\delta n_i = +1$
Birth of susceptible into rank i from parent in rank j	$r_b a_j \left(S_j + (1 - p_v) I_j\right) f_i \left(1 - \frac{N}{K}\right)$	$\delta S_i = +1$
Death in rank i	μn_i	$\delta n_i = -1$
Successful move from rank i into rank j	$\nu f_i n_i \frac{a_j}{N_l} e^{-\phi_p a_j n_j}$	$\delta n_i = -1$, $\delta n_j = +1$
Failed movement from rank i into rank j resulting in death	$\nu f_i n_i \frac{a_j}{N_l} \left(1 - e^{-\phi_p a_j n_j}\right)$	$\delta n_i = -1$
Birth of infective into rank j from parent in rank i	$r_b a_i p_v I_i f_j \left(1 - \frac{N}{K}\right)$	$\delta I_j = +1$
Death of infective in rank i	μI_i	$\delta I_i = -1$
Successful attempt to move by infective from rank i into rank j	$\nu f_i I_i \frac{a_j}{N_l} e^{-\phi_p a_j n_j}$	$\delta I_i = -1$, $\delta I_j + 1$
Failed attempt to move by infective from rank i into rank j resulting in death	$\nu f_i I_i \frac{a_j}{N_l} \left(1 - e^{-\phi_p a_j n_j}\right)$	$\delta I_i = -1$
Infective mover from rank i causing infection in rank j	$\nu f_i I_i \frac{a_j}{N_l} \beta_p \frac{S_j}{n_j}$	$\delta S_j = -1$, $\delta I_j = +1$
Global infection of susceptibles in rank i by infective in rank j	$\Gamma(1 - \alpha) S_i I_j$	$\delta S_i = -1$, $\delta I_i = +1$
Within rank infection in rank i	$\Gamma \alpha S_i I_i$	$\delta S_i = -1$, $\delta I_i = +1$

Table 2.3 Event rates and associated changes in state-space for the non spatial model with disease and without survival

Symbol	Description
State-space variables	
N (N_v)	Total number of individuals in group (group v)
n_i (n_{vi})	Number of individuals in rank i (in group v)
S_i (S_{vi})	Number of susceptibles in rank i (in group v)
I_i (I_{vi})	Number of infecteds in rank i (in group v)
N_l	Number of ranks in each group
Demographic parameters	
r_b	*Per capita* birth rate
K	Group carrying capacity
μ	*Per capita* mortality rate
a_i	Coefficient reflecting reproductive fitness and desirability of rank i
f_i	Coefficient reflecting probability of newborns entering and individuals leaving rank i
c	Intercept used in linear definition of a_i
ν	*Per capita* movement rate
p	Proportion of movements assigned to local movements
ϕ_p	Promotion constant: determines probability of success of entering a rank
μ_p	Probability of dying after failed movement attempt
ϕ_d	Dispersal constant: determines probability of success of entering a group
Disease parameters	
β_0	*Per capita* rate of group-level infection
β_1	*Per capita* rate of infection within rank
Γ	Overall strength of infection ($\beta_0 + \beta_1$)
α	Relative importance of group-level infection ($\beta_0/(\beta_0 + \beta_1)$)
p_v	Probability of disease transmission from parent to offspring at birth
β_p	Probability of disease transmission trough contact at movement attempt

Table 2.4 Definition of variables and parameters used in this chapter

Chapter 3

Parameterising wildlife population dynamics models using discrete observations with errors

3.1 Introduction

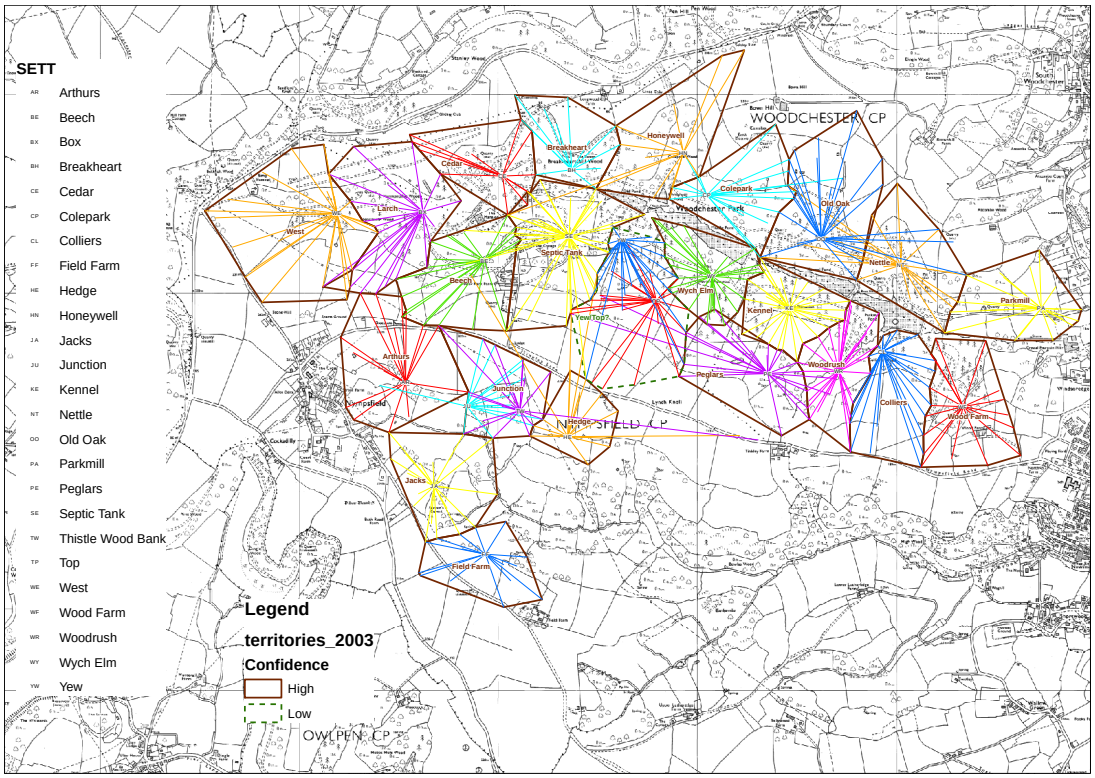
Understanding population dynamics processes in wildlife populations has important implications for wildlife conservation and management. It can also be an important component in understanding the spread of disease because wildlife species often live in small groups which can fluctuate significantly in size and this directly affects the disease dynamics within such populations (Anderson and May, 1979; Wilson and Hassell, 1997). Our understanding of demographic processes in wildlife populations is generally based on observations that are in some way incomplete. For example such observations are typically only made at discrete points in time (e.g. annually) and tend to include significant errors (depending on the sampling effort). Modelling can provide a way to deal with such shortcomings and help improve our understanding of demographic processes as illustrated by a wealth of modelling literature (among many others Anderson and Trewhella (1985) and Renshaw (1991)). However, the value of modelling greatly improves if it is integrated with the available data in a way that includes all sources of variability (e.g. demographic stochasticity, variations between locations) and uncertainty (e.g. observational error). Here we adopt an approach to inference in stochastic models documented by Marion et al. (2012) in the context of species distribution modelling. The focus of our application is the inference of demographic and (in Chapter 4) epidemiological processes in wildlife populations using discrete state-space continuous time Markov processes. In this context we introduce a number of novel developments which extend the state of the art including inference of observation model parameters and assessment of between group variability. This methodology is developed and tested in order to apply it to long term observations of a population of European badgers (*Meles meles*) in Woodchester Park (SW England) (Delahay et al., 2000, 2001; Vicente et al., 2007) (Figure 3.1(a)

and 3.1(b). These data have been collected since the mid 1970s and a consistent observation strategy has been in place since the early 1980s . Several authors have presented models for demography and disease using parameter estimates based on these observations (Delahay et al., 2000). The earliest such models were deterministic (Anderson and Trewhella, 1985; Bentil and Murray, 1993; Barlow, 1996; Swinton et al., 1997). Later models are more complex and include stochasticity (White and Harris, 1995a; Smith et al., 1995; Shirley et al., 2003). To date such models for badger ecology and epidemiology have relied on summary statistics (e.g. number of lactating females, life expectancy) for their parameter estimates and have therefore not accounted for the uncertainty inherent in the data. Our aim is to present a statistically rigorous approach for stochastic models that can be used to estimate the parameters for a simple population dynamics model directly from the Woodchester Park data.

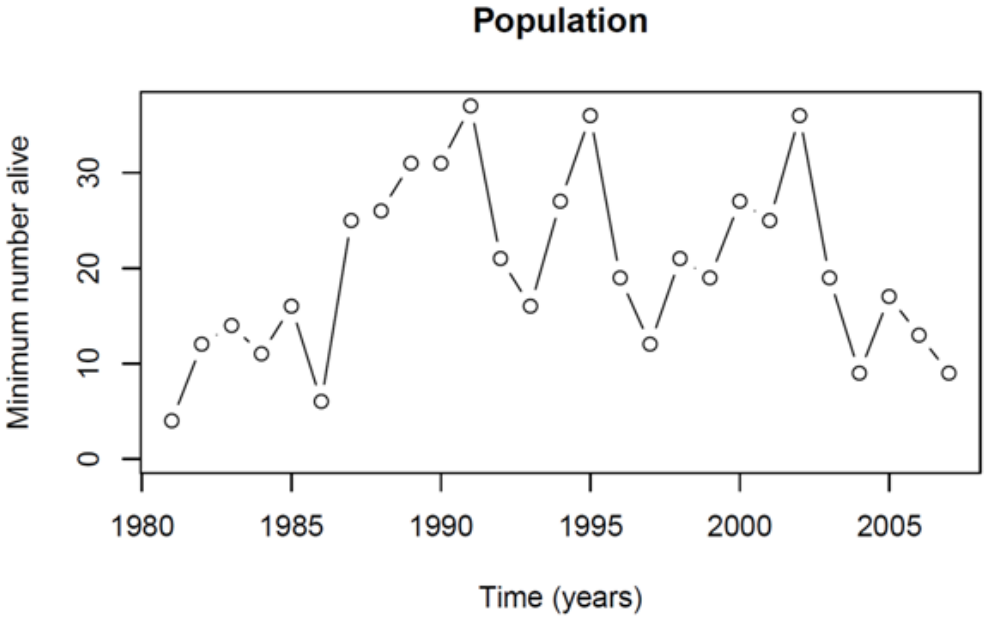
We implement Bayesian inference using Markov chain Monte Carlo (MCMC) to estimate the parameters for a *continuous time discrete state space model* for demography within a population affected by births and deaths only. We use an event based likelihood approach which was first proposed by Gibson and Renshaw (1998) and Renshaw and Gibson (1998). Their approach has subsequently been applied by various other authors including O'Neill and Roberts (1999), Gibson and Renshaw (2001), Marion et al. (2007), Boys et al. (2008) and Marion et al. (2012) to a variety of ecological and epidemiological systems. The method requires the construction of an event sequence that is consistent with the available observations of events and/or state space variables and allows joint estimation of parameters and missing data. Some authors have discussed applications to partially observed data, e.g. Gibson and Renshaw (1998), Renshaw and Gibson (1998), O'Neill and Roberts (1999) and Gibson and Renshaw (2001) discuss an application where all events of one or more types are observed, and all others need to be inferred. Others including Boys et al. (2008), Catterall et al. (2012) and Marion et al. (2012) apply the method to a case where only the state space variables and no actual events are observed. In none of these studies do the authors account for errors in the observation process but rather assume that any observations are exact. Moreover, the majority of such studies only consider fitting models to single replicate data sets. However, Gibson et al. (2006) analyse data on disease spread in plant populations from experimental microcosms in which they fit stochastic models to data from individual experimental replicates. Although these data exhibit considerable variability between replicates they make no allowance for this when fitting a single model to the combined data. Therefore methodological developments are needed to better account for observation errors and variability e.g. between groups in the context of wildlife populations. Note that there are several alternatives to the MCMC approach taken in this thesis, including Approcimate Bayesian Computation (ABC) and Particle MCMC.

The aims of the work presented here are:

1. To further develop the state of the art inference methodology using event-based likelihood approaches for continuous time stochastic models by (a) incorporating observation models, (b) combining the likelihood for multiple groups and (c) developing a hierarchical scheme to infer variation between



(a) Woodchester Park baitmarking map



(b) Example population history

Figure 3.1 (a) Baitmarking map for 2003 for the Woodchester Park study area outlining the territories of specific badger groups (Provided by Gavin Wilson (FERA), see Delahay et al. (2001) for details on baitmarking technique). For each of these groups we can extract a time-series of the population history. (b) An example for the WEST social group.

groups.

2. To develop and test the proposed inference methods for demographic processes in wildlife.

3. To demonstrate their practical utility by application to the long term observations from the Woodchester Park study.

When we study population dynamics in wild-life populations, we generally do not observe birth and death events directly and the population size is observed at discrete points in time (e.g. annually) via some imperfect sampling procedure. To date, when applying model parametrisation to this type of data the assumption has been made that the state space variables are observed without error. This can lead to considerable complications when constructing and proposing changes to the event sequence, for example Boys et al. (2008) implemented a complicated set of changes between observations in order to fit the event sequence exactly to the data. This type of complication does not occur when the data we use consists of events (e.g Gibson and Renshaw (1998) and O'Neill and Roberts (1999). As suggested by Marion et al. (2012) we can account for observational error by multiplying the event based likelihood with an *observation model* that relates the state-space observations to the inferred underlying state of the system. This removes the necessity of matching the event sequence exactly to the observations. In most previous work the observation model is not explicitly included (e.g. Catterall et al. (2012)) and thus they implicitly assume the the observations are made without error. Other authors have included an observation model but with fixed parameters assuming they are known (e.g. McInerny and Purves (2011). In our implementation we make the observation model and its parameters an integral part of the likelihood definition. This allows for more flexibility in the estimation and in theory allows us to estimate the noise parameters as well. The observation model specifies a probability distribution for any observed quantities conditional on the "true" underlying state of the system. Such observation models can take any number of forms depending on the type of limitations inherent in the observations (Royle and Dorazio, 2008; Marion et al., 2012).

The way in which we deal with the error term by including separate models for the process as well as the observations, is very similar to the approach routinely used in Bayesian inference on *discrete time state-space models* (e.g. Clark and Bjørnstad (2004), Buckland et al. (2007) and Newman et al. (2009)) and seems particularly suited to deal with capture mark recapture data (King et al., 2009). Although much progress has been made in developing inferential tools for such discrete time models, Buckland et al. (2007) have noted that continuous time state space models would in many cases provide a better representation of the underlying processes. In this thesis we demonstrate that we can overcome computational issues in order to implement inference for latent process models in continuous time.

3.2 Methodology

3.2.1 Generic inference approach

We start by outlining a generic framework for Bayesian inference in continuous time discrete state-space models. This follows, with slight modifications, the outline given by Marion et al. (2012) (Figure 3.2). Central to the approach is the likelihood which can be defined as the probability of making observations D given a model structure, parameter values and some initial condition. The model definition can be split into (a) the process model which describes the underlying biological assumptions and can be thought to represent the "true" but unobserved state of the system of interests, and (b) the observation model which describes the probability of making the observations D for any underlying state of the system as described by the process model.

Understanding of the biological processes that determine the system of interest enables specification of the stochastic process model $P(\zeta|\theta_M, s(t_0))$ of any complete state-space history ζ (e.g. the population size at all times of interest). The process model is conditional on the model parameters θ_M and the initial state $s(t_0)$ (e.g. the initial population size). The observation model $P(D|\zeta, \theta_N)$ describes the probability of observing the data D given any underlying history ζ. The product of the process model and the observation model is the complete likelihood for the parameters $\theta = (\theta_M, \theta_N, s(t_0))$. Bayesian inference is based on the posterior distribution of the unknowns (θ, ζ) given the data, which is proportional to the product of the prior and the complete likelihood. The prior $P(\theta)$ encodes knowledge (sometimes obtained directly from alternative data sets) that constrains the range of parameter values. In practice, techniques such as MCMC are used to draw samples from the posterior, which in turn can be used to obtain almost any relevant statistic. For example parameter distributions $P(\theta|D)$ can be obtained by marginalizing over possible histories ζ.

3.2.2 Definition of the event based process likelihood

We follow the definition of an event based likelihood for continuous time state-space models proposed by Gibson and Renshaw (1998) and the notation of Marion et al. (2007). In the case where all changes to the state-space (events) are observed the complete state space history ζ is fully described by the initial state $s(t_0)$ and the set of all events that occur over the same time period $E = \{(E_k, t_k) : k = 1 \dots n\}$, where $E_K \in \{e_i : i = 1 \dots \#event\ types\}$ and e_i is the set of possible event types, and the time at which events occur $t_k \in [t_0, t_n]$. The process model, in this case referred to as the *complete likelihood*, for this complete state space history ζ and the process parameters θ_M and the initial population $N(t_0)$ is

$$P(\zeta|\theta_M, N(t_0)) \propto \prod_{k=1}^{n} P(E_k|N(t_{k-1}); \theta) \tag{3.1}$$

where the probability of a specific event E_k occurring at t_k is defined as

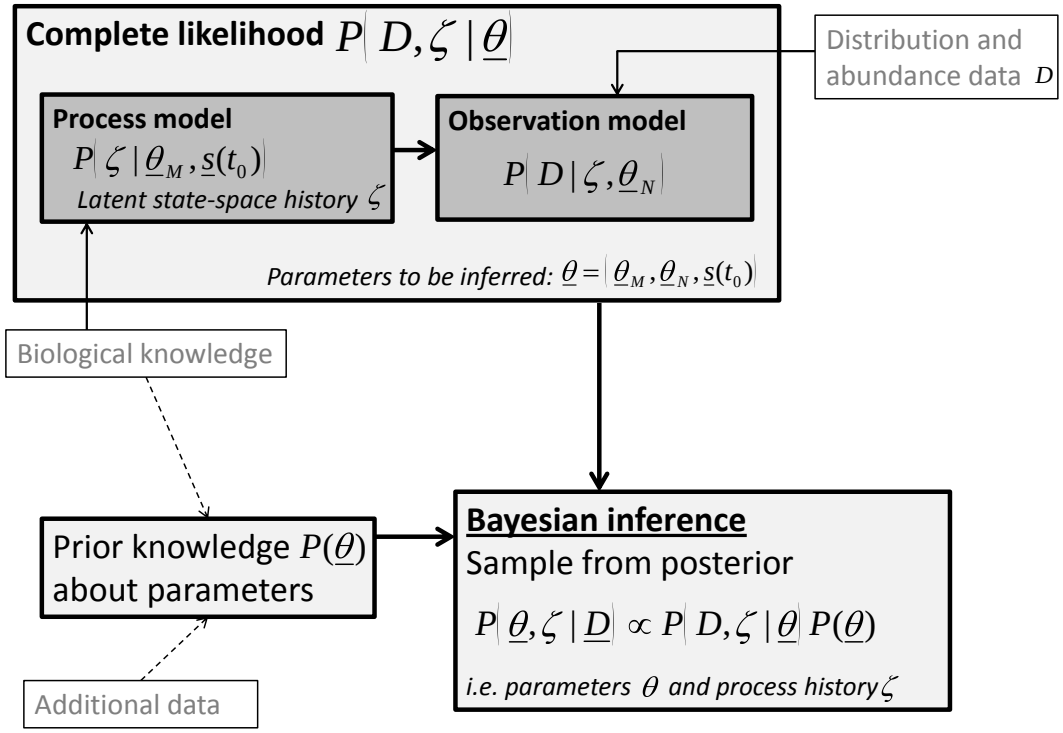


Figure 3.2 Generic overview of inference approach for dynamic stochastic models. It includes model components in dark grey boxes, inference elements light grey and biological data and information white boxes. See text for details. Modified from Marion et al. (2012).

$$P(E_k|N(t_{k-1});\theta) \propto r(e_i,N(t_{k-1});\theta_M)e^{-R(N(t_{k-1});\theta_M)\delta t_k} \tag{3.2}$$

where $R(N(t);\theta_M)$ is the sum of all the event rates at time t. In the case of our simple model for population dynamics we only need to consider birth and death events. To account for the fact that the end of the period of interest t_n does not necessarily coincide with an event, the likelihood also includes the probability of no event happening between the last event and t_n. Having defined the likelihood function, in the case where we have observations of all events, we could use a maximum likelihood method to find the parameter estimates Edwards (1992) but here we proceed with a Bayesian approach.

3.2.3 The observation model

When dealing with ecological observations we rarely have complete knowledge of the timing and order of events. In our particular application for the long term observations of badgers in Woodchester Park (Section 3.4) we only use an annual count of the number of badgers and these counts are subject to considerable error as only a proportion of all individuals is captured every year. In these

circumstances, the initial state and the timing and type of any events that are not observed need to be inferred in such a way that the corresponding inferred state space sequence is consistent with the observations available as well as the process and observation models. We think of the complete set of events as $E = \{D_E, H\}$, where the sub-set D_E is the set of all observed events and the set H is the set of all unobserved events (sometimes referred to as hidden). The set of all data D includes observations of the state-space D_S as well as observed events D_E, so that $D = \{D_E, D_S\}$. In our case we have no observations of events (e.g. births and deaths) so the set D_E is empty. However, in some systems events may be observed directly (e.g. animal movements (Walker et al., 2006)).

In the estimation procedure, the details of which will be discussed later, we infer the state space history. These, together with the initial population (which is also inferred) define an inferred history of the state-space S. We relate the observations of the state-space D_S to the inferred history of the state space S through an observation model $P(D_S|S, \theta_N)$, where θ_N is a set of observation model parameters. When generating simulated data and in the application to the Woodchester Park data the observation model is the binomial probability of observing a particular population size given the "true", inferred population size and some probability of detection p_d for each individual. In other applications the observation model can take other forms depending on the system of interest and the way it is observed (Marion et al., 2012), for example a Bernouilli distribution for presence absence data or the assumption that observations are exact (Catterall et al., 2012).

The combined likelihood of the data and the hidden events is given by

$$P(D, \zeta|\theta) = P(\zeta|\theta_M, s(t_0))P(D|\zeta, \theta_N) \tag{3.3}$$

and this is used in Bayes's Theorem to define the posterior distribution for the parameters θ (including both model and noise parameters) and the missing the unobserved state space history ζ as

$$P(\theta, \zeta|D) = \frac{P(D, \zeta|\theta)P(\theta)}{P(D)}. \tag{3.4}$$

where $P(\theta)$, assuming the parameters are a priori independent, is the product of the prior probabilities of all parameters and $P(D)$ is constant for a given data-set. So for a given data-set D we have

$$P(\theta, \zeta|D) \propto P(D, \zeta|\theta)P(\theta) \tag{3.5}$$

.

3.2.4 Sampling parameters from the posterior

Markov chain Monte Carlo techniques allow us to draw samples of the parameters θ and state space histories $\zeta = H$ from the posterior $P(\theta, \zeta|D)$ without having to calculate the normalisation constant $P(D)$. In this approach we generate successive samples of θ and ζ using a Markov chain which should converge to

a dynamic equilibrium that represents the posterior. The first samples before reaching this equilibrium need to be discarded (this is commonly referred to as the burn-in period) and we may need a large number of samples before achieving convergence. Additional information about the implementation of the inference algorithm is given in Appendix D.2.

We can use the Metropolis Hastings (MH) sampler (Metropolis et al., 1953; Hastings, 1970) in which we make small changes to the current estimate of the parameter value and accept or reject the new value depending on the ratio of the likelihoods calculated with the current and proposed parameter value. Dependent on the model structure and for some choices of parameter and prior distribution it is possible to calculate the exact form of the marginal posterior. This then enables parameter samples to be drawn directly from the marginal posterior - a procedure known as Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990). In many cases this is more efficient than the MH algorithm because no samples are rejected.

We can use the marginal posterior for r_b as an example of a marginal posterior that can be written as one of the standard distributions. Starting with the full posterior for our model

$$P(\theta_M|\zeta) = \frac{P(\zeta|\theta_M)P(r_b)P(\mu)P(K)}{P(\zeta)} \tag{3.6}$$

and choosing the prior for r_b as

$$P(r_b) \sim \Gamma(\alpha_{r_b}, \beta_{r_b}) \propto r_b^{\alpha_{r_b}-1} \exp\left(-\beta_{r_b} r_b\right) \tag{3.7}$$

which is a gamma distribution with parameters α_{r_b} and β_{r_b}. Then the posterior density of r_b conditional on μ, K and ζ is also a gamma distribution

$$\begin{aligned} P(r_b|\theta_{-r_b}, \zeta) &\propto r_b^{n_b+\alpha_{r_b}-1} \exp\left(-r_b(\beta_{r_b} + \beta'_{r_b})\right) \\ &\sim \Gamma(\alpha_{r_b} + n_b, \beta_{r_b} + \beta'_{r_b}) \end{aligned} \tag{3.8}$$

where n_b is the number of birth events in the state space history and using the definition of our process model

$$\beta'_{r_b} = (t_f - t_n)N_n\left(1 - \frac{N_n}{K}\right) + \sum_{k=1}^{n}(t_k - t_{k-1})N_{k-1}\left(1 - \frac{N_{k-1}}{K}\right). \tag{3.9}$$

so to sample from the marginal posterior we simply keep track of the number of birth events in the state space history and recalculate β' in each iteration. The same holds for any other parameter (in our case this includes the death rate μ) as long as the rates of the Markov process used in the likelihood are linear in that parameter and we choose an appropriate form for the prior (e.g. gamma or beta).

Cowles and Carlin (1996) discuss a range of heuristic tests of convergence for MCMC, but in general there are no exact methods to determine convergence. In this study we mostly use qualitative criteria to judge whether an inference run

has converged to the stationary distribution. In all cases we visually check the trace plot (e.g. Figure 3.4) for convergence and discard the samples generated before convergence was reached (burn-in). We assume that convergence has occurred once the trace plot fluctuates around some "stable" equilibrium that is maintained over a long period. When working with simulated data, a good estimate is considered to be one where the "true" parameter value used to generate the simulated data is included in the 95% Credible Interval. In some of the early results (Section 3.3) 3.3 we also report the coverage which is the percentage of a large number of runs for which the 95% CI includes the "true" value, and the percentages of the number of runs which have median above or below the "true" value which can be seen as a measure of the balance of the estimate. In later sections we run the inference on a smaller number of groups and report the median and 95% CI and plot the posterior density distributions which give a more complete impression of shape of the distributions. In general we assume that a more focused posterior (i.e. smaller variance) is better than a wider posterior provided it covers the "true" value. This allows us to compare estimates that are derived in different ways (single group estimates versus combined likelihood estimates based on multiple groups). A final criterium for the performance of an inference run is whether the resulting estimate is robust when using different and more challenging prior distributions. Obviously many of the qualitative criteria mentioned here can not be used when interpreting results using real data (Section 3.4). In that case we assume that the results can be trusted if we can demonstrate that the approach works on simulated data produced with similar parameter values.

3.2.5 Sampling state space histories from the posterior

When sampling state space histories we must explore spaces of varying dimensions, therefore we use reversible jump MCMC (Green, 1995), an extension of the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), to generate samples of the unobserved state space history ζ from the posterior $P(\theta, \zeta|D)$ (Gibson and Renshaw, 1998). We start off with an initial state, event sequence and corresponding state space history that are consistent with the data. During each iteration in the inference run we propose a small number of changes $M > 0$ to the event sequence. The new event sequences resulting from each change is accepted with probability

$$A_H = min\{1, LR \times NR \times PR\} \tag{3.10}$$

where

$$LR = \frac{P(\zeta'|\theta_M)}{P(\zeta|\theta_M)} \tag{3.11}$$

is the likelihood ratio in which ζ is the current state space history and ζ' is the proposed new state space history,

$$NR = \frac{P(D|\zeta', \theta_N)}{P(D|\zeta, \theta_N)} \tag{3.12}$$

is the ratio of the observation model terms in which ζ is the current state space history and ζ' is the proposed state space history, and

$$PR = \frac{q(H' \to H)}{q(H \to H')} \tag{3.13}$$

is the proposal probability ratio, where $q(H \to H')$ is the probability of changing the hidden part of the event sequence H to H' and $q(H \to H')$ is the probability of reversing that change.

There are many options for initialising the event sequence but in our application we start off with no events and the initial population $N(t_0)$ set to the maximum observed population size. Possible changes are the addition of new events (i.e. births or deaths), the removal of existing events, changing the timing of existing events and changing the initial state (here we are only concerned with population size). Note that some proposed changes may result in "impossible" state space histories (e.g. changing the intial population to negative values, changes that result in population sizes smaller than the carrying capacity). In that case the likelihood is set to a negative value which automatically leads to acceptance probability of zero. Changes to the event sequence are referred to as moves and are proposed with a specific proposal probability. Some care is needed to work out the proposal probability for each move:

1. Addition of an event (birth or death) with probability p_1. In this case we draw a unique random time to insert an event from $U(t_0, t_n)$, then we work out the type of event to add with probability

$$p_{add} = p_1 \times p(type)\frac{1}{T} \tag{3.14}$$

   where $T = t_n - t_0$.

2. Removal of an event with probability p_2. In practice we select a random event from the event sequence with probability $1/n$, so the proposal probability for removing an event in this way is:

$$p_{del} = p_2\frac{1}{n} \tag{3.15}$$

3. The change of the timing of a randomly chosen event with probability p_3. Here we randomly choose one event with probability $1/n$, and draw a new unique random time from $U(t_0, t_n)$ for that event. The proposal probability

for changing an event time in this way is:

$$p_{shift} = p_3 \frac{1}{n \times T} \tag{3.16}$$

4. The change of the initial population $N(t_0)$ with probability p_4. Here we randomly choose whether to increment or decrement the initial population. This is done with equal probability so the proposal probability for this type of change is simply:

$$p_{init} = p_4 \frac{1}{2} \tag{3.17}$$

The proposal probability ratio is then given by

$$PR = \begin{cases} \dfrac{p_2}{p_1 \times p(type)} \dfrac{T}{1} \dfrac{1}{n+1} & \text{add an event} \\ \dfrac{p_1 \times p(type)}{p_2} \dfrac{n}{1} \dfrac{1}{T} & \text{remove an event} \\ 1 & \text{change timing of an event} \\ 1 & \text{change the initial population} \end{cases} \tag{3.18}$$

3.2.6 Prior distributions

The prior probabilities of the parameters should reflect what we know about the parameters before including the data D (Kass and Wasserman, 1996). This could be based on the analysis of data from other sources or on expert knowledge about the processes involved. In principle prior distributions can take many different forms, including *vague* priors expressing that we only have vague or general knowledge about a particular parameter (e.g. uniform prior over the entire possible range of a parameter). Such priors are often referred to as *uninformative* but it has been argued by Irony and Singpurwalla (1997) that truly uninformative priors do not exist. A priori we assume the parameters are independent and where possible we choose prior distributions that results in the most convenient sampling procedure.

In our application, choosing a Gamma distribution for the birth rate (r_b) and the death rate (μ) results in marginal posteriors that are also Gamma distributions from which we can directly sample using a Gibbs algorithm. There is no convenient choice for the carrying capacity K for which we implemented a Metropolis Hastings sampler.

In our approach we do not fix the parameters for the observation model and in theory these can be estimated as well. In practice, we find that the type of data we use in this study does not contain enough information to do this (e.g Section 3.3.4. The estimate returned is often not very different from to the prior used for the parameters for the observation model. In practice however, allowing for some

flexibility in the observation parameters seems to result in better convergence for the other parameters. In our application , choosing a Beta distribution for the detection probability p_d results in a marginal posterior which is also a Beta distribution, which can be sampled with a Gibbs sampler. However when we extend the model to include disease (Chapter 4), the error model includes a the product of the detection probability and the test sensitivities. In that case the marginal posterior can not be rewritten as a standard distribution. For this reason we have implemented the sampling of p_d as a Metropolis-Hastings sampler.

3.3 Application to simulated data

We present a simple model for population dynamics that defines the rates in the event based process likelihood and can be used to simulate data that we can use to test the inference approach. The model is designed to be applicable to population dynamics in badgers and the data we simulated is aimed to reflect the type of data we have available from Woodchester Park (an example is given in 3.1(b). Using the simulated data we demonstrate that our inference approach can be used to derive meaningful estimates of the model parameters before we apply it to the data from Woodchester Park (Section 3.4).

3.3.1 A simple model for population dynamics

Population dynamics can be modelled stochastically by treating them as a discrete state-space continuous time Markov process with exponential waiting times as outlined by Cox and Miller (1965) and Gillespie (1977). In this approach we define the rate at which each possible type of event e_i occurs as a function of the state $s(t)$ of the system and the set of parameters θ_M that govern the process. We adopt a simple model for population size $N(t)$ at time t which is only affected by births (e_b) and deaths (e_d). Births occur at logistic rate $r(e_b|N(t);\theta) = r_b N(t)(1 - N(t)/K)$ and deaths at rate $r(e_d|N(t);\theta_M) = \mu N(t)$, where r_b is the *per capita* birth rate, K is the maximum possible population size (often referred to as the *carrying capacity*) and μ is the *per capita* death rate. Note that here we only include a single state-space variable (i.e. $s(t) = N(t)$) and the set of model parameters is $\theta_M = \{r_b, \mu, K\}$.

In order to test the implementation of the inference procedure we ran a number of simulations for a population affected by births and deaths only (Figure 3.3). Two different observation procedures were applied to each run. One where the population size was recorded without error as and when events occurred (red line in Figure 3.3. This data set is referred to as the complete data-set or the complete set of events. The simulations also includes the option to observe the population at discrete points in time and individuals are counted with a given binomial detection probability p_d (blue crosses in Figure 3.3). This *observation model* is identical to the observation model in the inference code and this data set is referred to as the discrete or incomplete data set. The latter type of data is very similar to the data we extract from the Woodchester Park database (Figure 3.1(b)) to which we ultimately apply the inference method (Section 3.4). The

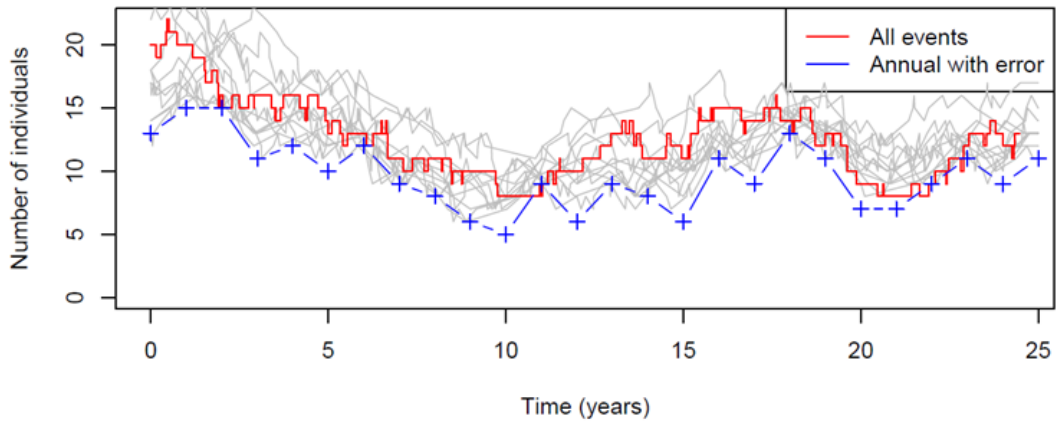


Figure 3.3 Example of a simulated complete event sequence (red) simulated using birth rate $r_b = 0.4$, death rate $\mu = 0.2$ and carrying capacity $K = 40$. Every individual present at annual observation points was observed with a detection probability ($p_d = 0.8$). The grey lines are 10 examples of event sequences that where accepted during the course of an inference run.

values of the parameters used to simulate the data were taken from the modelling literature for badgers and TB (*Mycobacterium bovis*), a discussion of which will follow later (Sections 3.4 and 3.5).

3.3.2 Inference on complete data

Although we rarely have complete data available we can use the complete dataset to illustrate and check the inference method. It may also help us to assess limitations in our abilities to infer process parameters and for example whether whether we can expect the method to work equally well for all parameters. Figures 3.4 and 3.5 illustrate that with the complete data set the method is highly successful. The priors we used to obtain these estimates were set to be weak (grey lines in density plots in Figure 3.4) and the parameter estimates seem very well constrained (black lines in density plots). If we apply the inference on 100 independently simulated data sets representing different groups the "true" value for all parameters is contained in the 95% Credible Interval for at least 96 of the data sets (Figure 3.5). There appears to be a tendency to overestimate the carrying capacity K (29% of the median estimates are below the "true" value and 71% above). A possible explanation for this is that the estimate of the carrying capacity for a particular group is bounded below by the maximum population size within that group. In a sense the likelihood is cut-off at that population size (i.e. $P(K < N_{max}|D) = 0$) and lower values of the carrying capacity are therefore automatically rejected. Note that it might be possible to to reduce the tendency to overestimate K by using an asymmetric prior, but from a biological point of view the carrying capacity is not very well defined. We tend to have a better understanding of the realised (equilibrium) population size which also depends on the birth and death rates. This suggests that a better alternative may be to constrain the estimates of K by using stronger priors for r_b and μ. The other

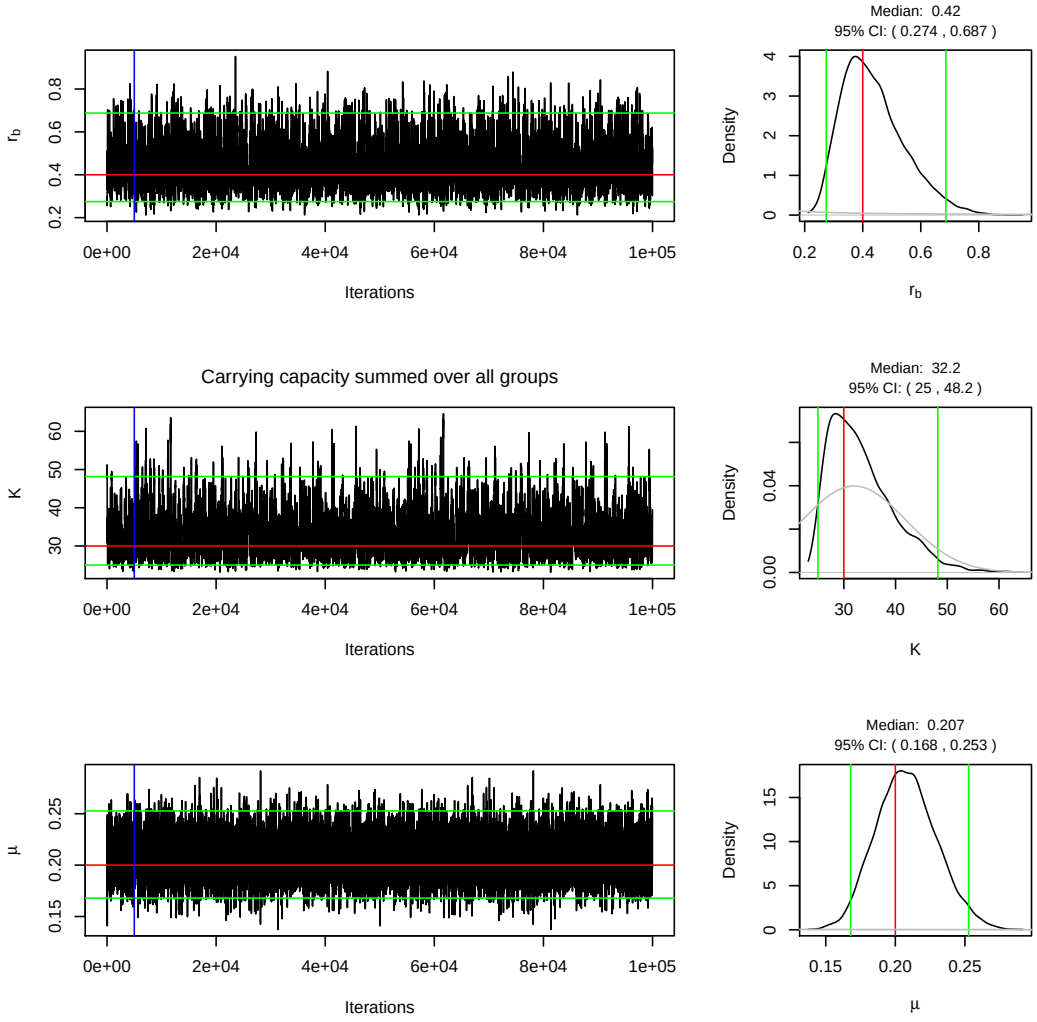


Figure 3.4 Example of inference results for demographic parameters (r_b, K and μ) using a complete data set (i.e. N observed every time an event occurs) from a simulation. For each parameter the sample is plotted as a trace on the left and the resulting posterior probability density on the right. These can compared with the "true" value of the parameter (red line) that was used to simulate the data, and with the prior density (grey line in density plots). The green lines indicate the boundaries of the 95% Credible Interval for each parameter. Above each density plot we also list the mean and standard deviation of the estimate. The samples from the "burn-in" period (5000 iterations, indicated by blue lines on the trace plots) are not included in the posterior density estimate.

parameters are not affected by such a lower bound as evidenced by the number of cases in which the median under or over estimates their "true" value.

3.3.3 Inference of events only on annually observed data

To allows us to assess how well the initial population, detection probability and unobserved birth and events are estimated we ran the inference on incomplete

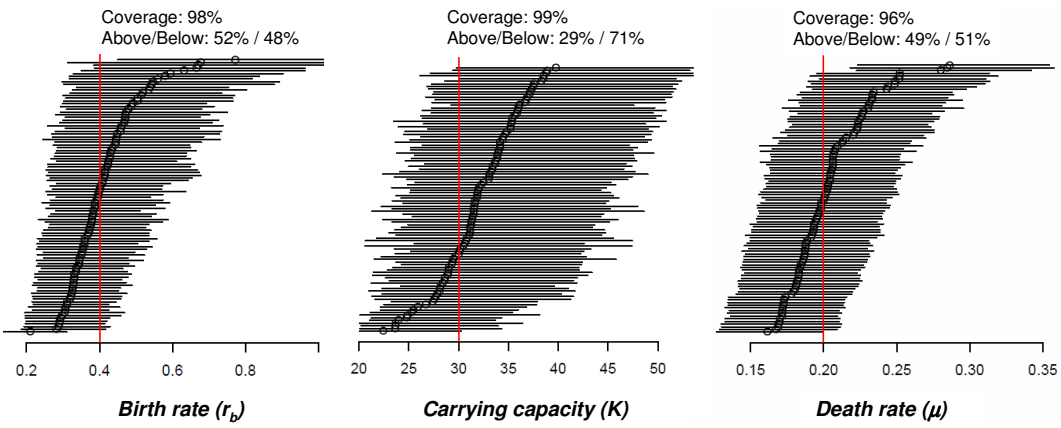


Figure 3.5 Summary of inference results for the 3 demographic parameters on 100 simulated complete data sets. Black dots represent the median of the posterior estimates and the bars span the 95% Credible Interval(green lines in Figure 3.4). The red line indicates the "true" value of each parameter used to simulated the data. Above each plot the percentage of cases where the 95%CI includes the "true" parameter value as well as the percentage where the median was below or above the "true" value respectively are indicated as well.

data with the parameter values fixed at their "true" values. In Figure 3.6 we show the results for one particular simulated data-set. In this example the inference was run over 10000 iterations with $M = 50$ proposed changes in each, so in effect we generated 50000 samples of possible of the state space history (10 examples of such state space histories are given in Figure 3.3 (grey lines)). Figure 3.6(a) shows that the detection probability p_d is estimated well in this scenario where the birth and death parameters were not inferred but fixed at the values used to generate the data. We also register the initial population (Figure 3.6(b) and the number of births and deaths (Figures 3.6(c) and (d)) in each of the accepted event sequences throughout the inference run. The resulting 95% Credible Intervals in the example in Figure 3.6 all cover the "true" numbers in the complete data-set from which the annual observations were generated. When we repeat this for 100 simulated data-sets we cover the "true" initial population in 98% of the cases and the "true" number of births and deaths in all of the cases.

3.3.4 Inference of parameters and events on annually observed data

Our main interest is in estimating the parameter values for our simple model for population dynamics from incomplete data. Figure 3.7 shows the results for 50 simulated data-sets that were annually observed. We compare the posterior probability density for individual runs (grey) with a *combined single-group estimate* (blue). The latter is derived by averaging the density of the 50 estimates and

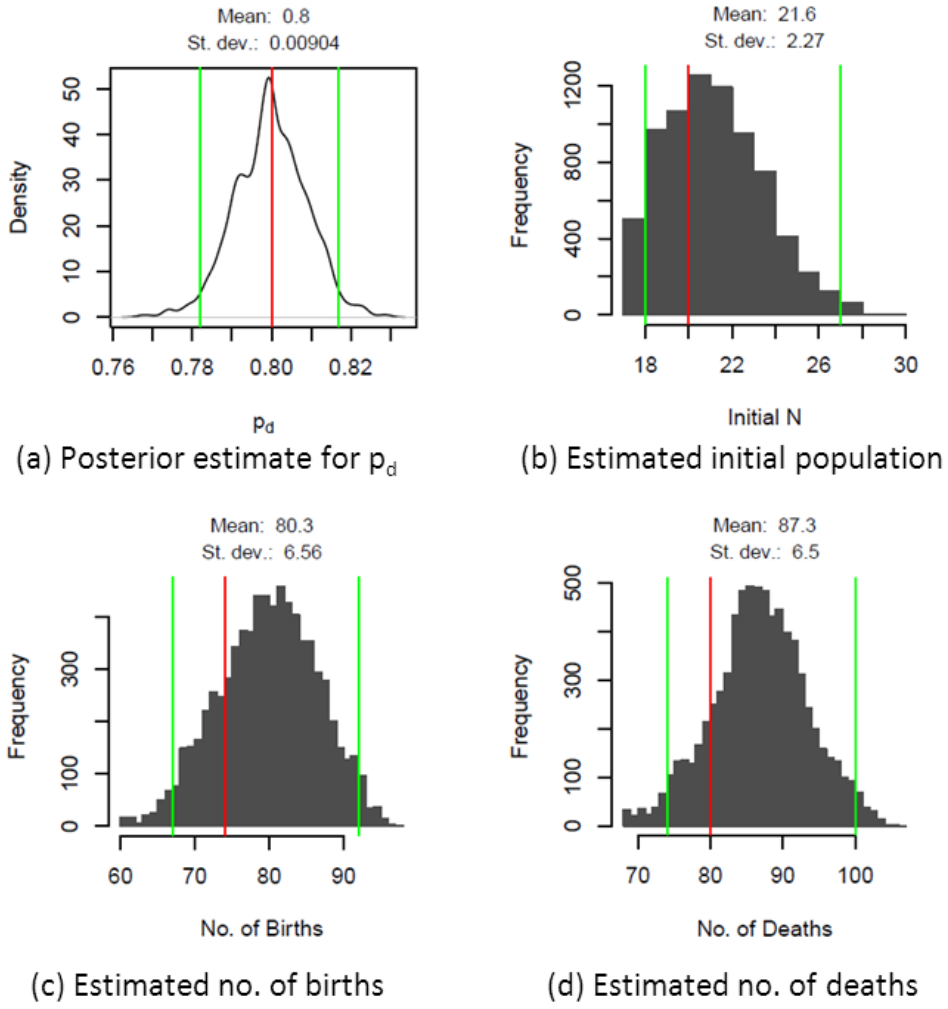


Figure 3.6 Estimates of the detection probability p_d (a), initial population N(b) and numbers of births and deaths (c,d) in inferred event sequences from one inference run on annually observed data. The "true" values in corresponding complete event sequences are plotted in red, and green lines show the boundaries of the 95% CI. These results were generated with birth and death rates fixed at their "true" values. Above each plot we also list the mean and standard deviation of the estimate.

is a crude way to summarise the uncertainty expressed by by the single group estimates. Note that other authors have used to product of estimates from single experiments (Gibson and Renshaw, 1998) but such estimates tend to be dominated by the prior and do not reflect the variation between replicates as seen in Figure 3.7. In section 3.3.5 we discuss a more rigorous approach to incorporating data from multiple replicates. The plots in Figure 3.7 illustrate that there is good correspondence between the median of the combined single-group estimate and the "true" parameter values. However the combined single-group estimates (average of densities for all groups) for all parameters apart from the death rate μ are very close to the prior distributions (green shading) and thus the data does

not appear to provide further constraint on the parameter estimates.

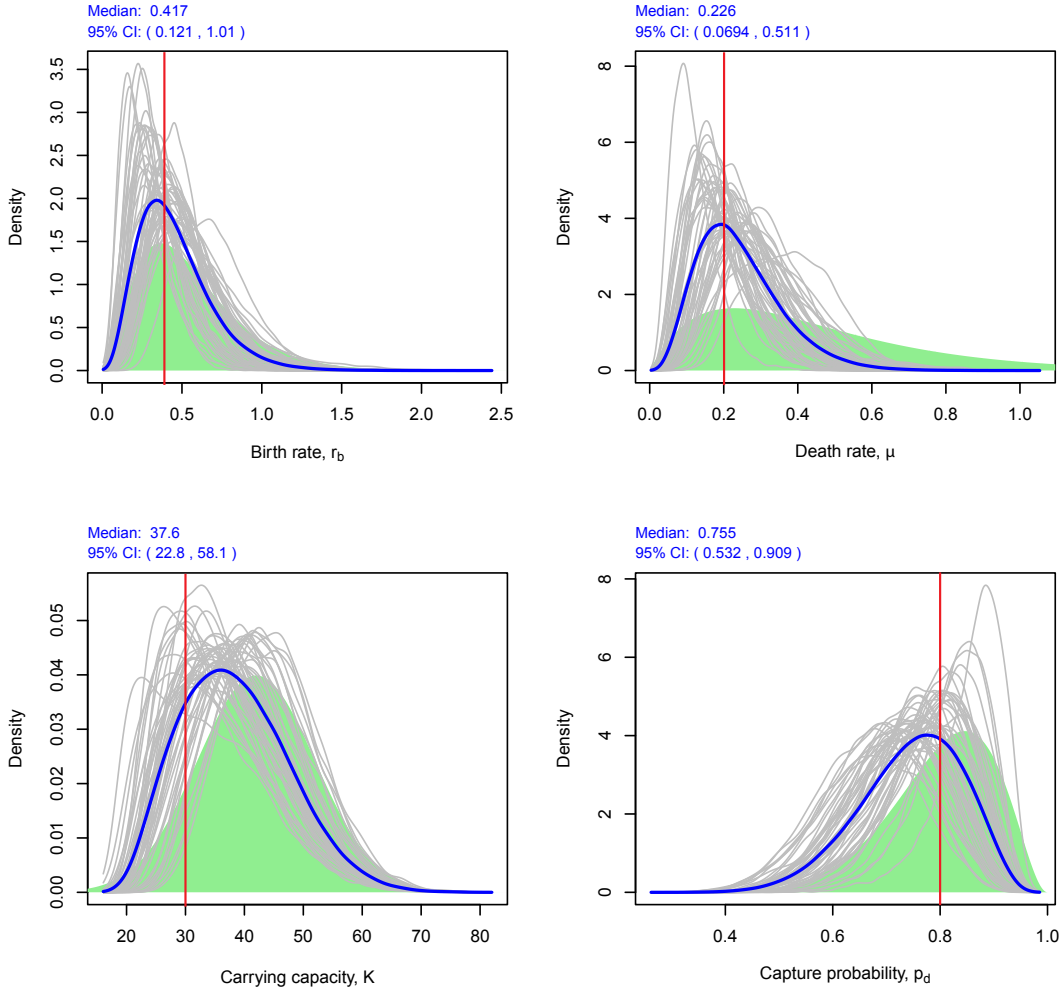


Figure 3.7 Posterior probability densities (grey) for each of the demographic parameters for 50 simulated data sets that were observed annually. The blue line marks the combined single-group estimate which is the average density of all fifty single-group estimates and the green shading indicates the prior. The "true" parameter values, used to generate the simulated data are indicated in red. Above each density plot we also list the median and 95% Credible Interval of the estimate.

Improved estimates can be obtained by using more frequently observed version of the simulated data (Figure 3.8). If the population size is recorded every half year or every month instead of annually, the combined single-group estimates for the birth and death rates become more focused (tighter). The estimate for the carrying capacity does not appear to improve. These estimates are for a restricted uniform prior for the detection probability ($p_d \sim U(0.61, 0.99)$). Interestingly, the estimates of p_d based on annual and half yearly observations seem to favour the lower half of the prior interval. The estimates of p_d using monthly estimates are evidently much better. This shows that in theory if we have sufficient data we can estimate the parameters of the observation model in conjunction with the

process parameters. Unfortunately it is not feasible to extract observations at a higher than annual frequency from the Woodchester Park database (Section 3.4) so we can not rely on this approach to improve the parameter estimates.

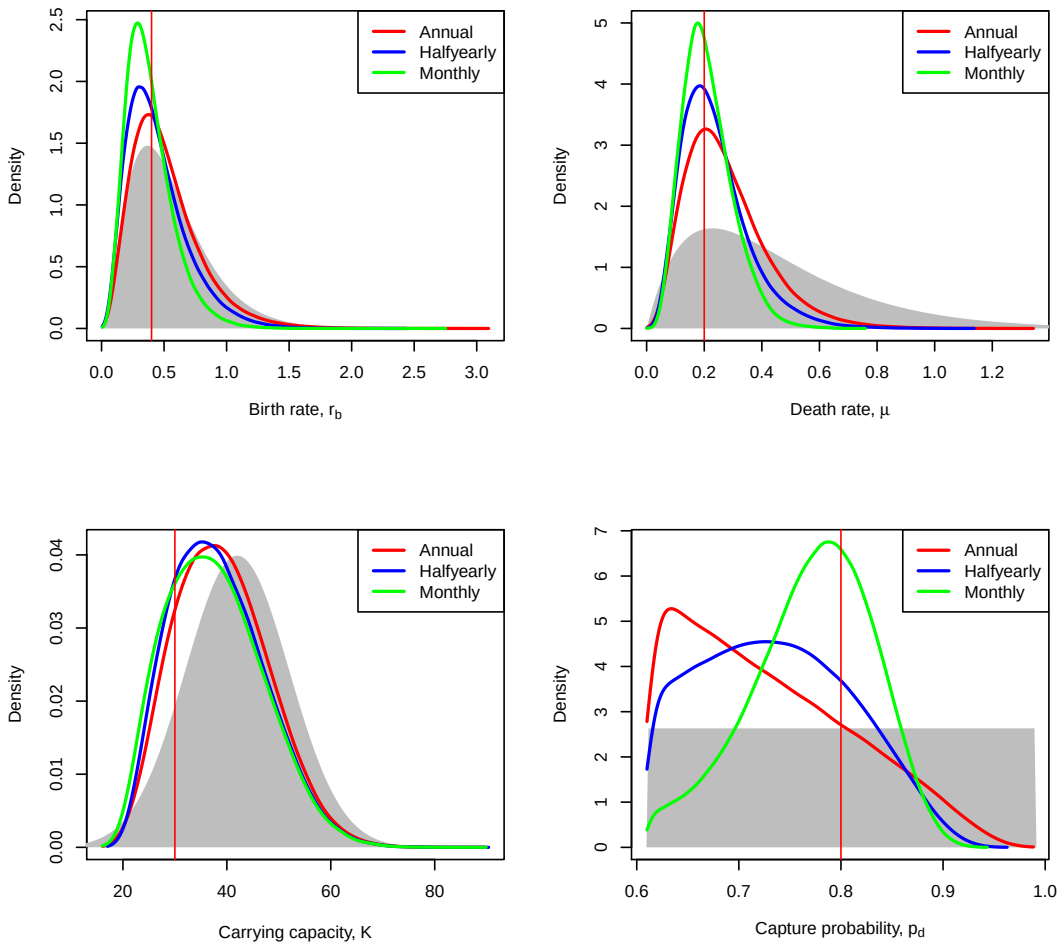


Figure 3.8 Posterior probability density estimates for each of the demographic parameters generated with simulated data-sets with varying observation frequencies (annually, half-yearly and monthly). Vertical red lines mark the "true" parameter values and the grey shadings mark the prior for each parameter. Note that we used a uniform prior for the detection probability ($p_d \sim U(0.61, 0.99)$).

3.3.5 Combining the likelihood of multiple groups

In the Woodchester Park database we have information about the number of individuals present in about 25 distinct social groups (Figure 3.1(a)). Although there is some interaction between groups of badgers in Woodchester Park (migration, transmission of disease between individuals in different groups), these are processes are thought to only play a minor role. So in order to keep the model simple we assumed that there is no interaction between groups and that the demographic processes in each group are governed by the same parameters. In doing that we can combine the likelihood of groups by simply multiplying

them to obtain a better constrained estimate of the parameters (Gibson et al., 2006). It is clear from Figure 3.9 that including more data in this way results in much better estimates of the three process parameters. The posterior probability densities based on the combined likelihood (red) are much tighter when compared to the estimate based on the combined single group estimates for single groups (blue and grey). The estimates based on the combined likelihood (red lines) all cover the "true" parameter values and display much less overlap with the prior distributions (green shading). Note that in the particular example show in Figure 3.9, the prior for the carrying capacity ($K \sim N(42, 10)$) is offset from the "true" value of $K = 30$ and the estimate is also offset in the same way. It was noted in Section 3.3.2 which discusses the use of complete data that the lowest possible estimate of K is the maximum population that occurs in the observations. Consequently, because we assume that the parameters are the same in all groups the lower bound of K is determined by the maximum population observed in all the groups included in the estimate.

Figure 3.9 also suggests that the estimates for K and p_d are linked with higher estimates of K resulting in lower estimates of p_d. A joint density plot for K and p_d (Figure 3.10) confirms this link, albeit fairly weak. It can can be understood by looking at the definition of our model. The deterministic equilibrium population for our model can be defined as $\tilde{N} = (1 - \mu/r_b)K$ and thus if we over-estimate K as argued in the previous paragraph (assuming the ratio between μ and r_b is estimated correctly) then this will be reflected in an overestimate of the population size N. Overestimating the inferred population size causes a larger difference between the observed and inferred population size and thus is reflected in an underestimate of p_d.

3.3.6 Understanding the effect of more data

Our results so far have shown that increasing the amount of data available for estimation either through increasing the observation frequency (Section 3.3.4, Figure 3.8) or by including data from multiple simultaneously observed groups (Section 3.3.5, Figure 3.9) can have a marked effect on the quality of our estimates. In this section we will briefly explore the reasons for this by analysing the mathematical definition of the posterior.

Assume an arbitrary process parameter p which occurs as a linear factor in the definition of the Markovian transition rates that define the process model. Then if we use a prior $P(p) \sim \Gamma(\alpha, \beta)$ the posterior can be defined as

$$P(p|\zeta) \propto \Gamma(\alpha + n_p, \beta + \beta') \tag{3.19}$$

where n_p is the number of events of type e_p and

$$\beta' = \int_{t_0}^{t_f} f(s(t))\,\mathrm{d}t \tag{3.20}$$

where t_0 and t_f are the beginning and end of the period over which we study the process.

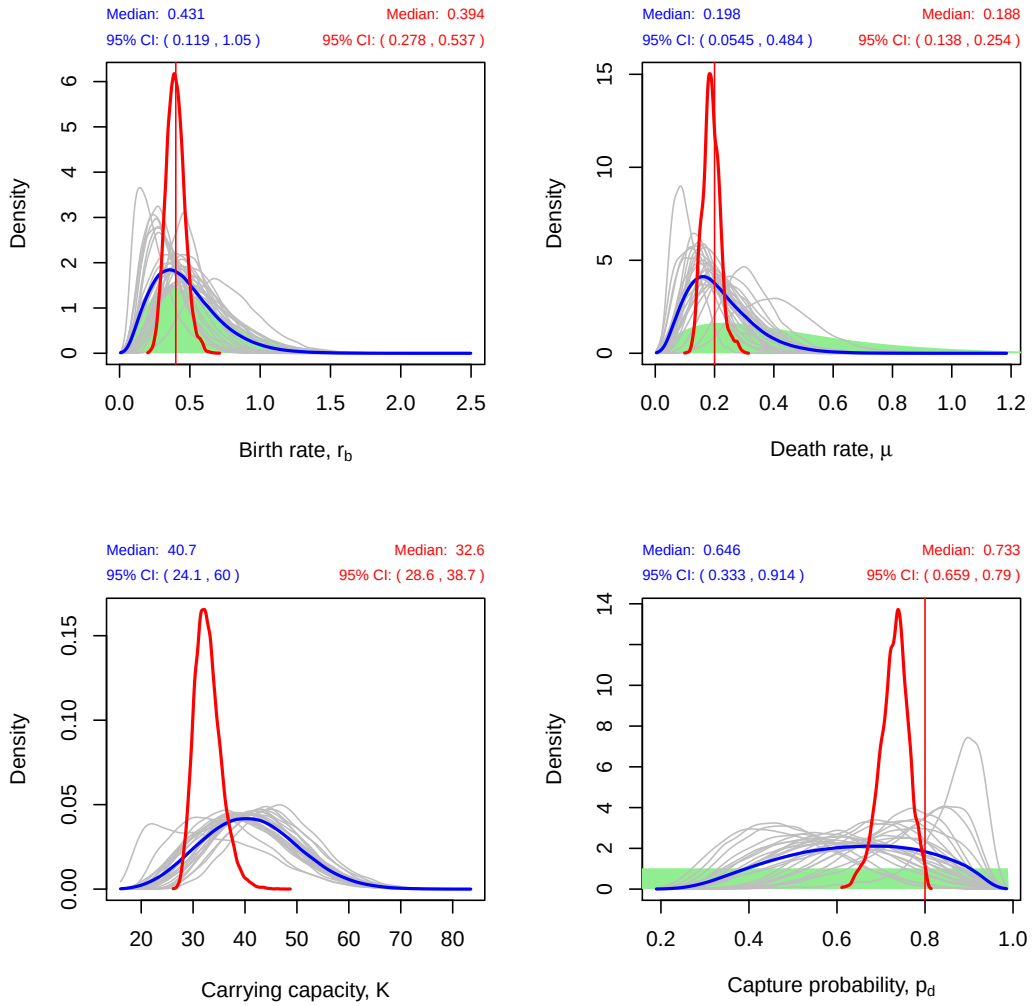


Figure 3.9 Posterior probability density estimate for each of the demographic parameters derived using a combined likelihood for 25 groups (red curve) on annually observed simulated data. Included for comparison, posterior probability densities for 25 simulated data sets (grey) and the combined single group estimate (blue, average of grey curves) of all 25 estimates. A burn-in period of 100000 iterations was applied to the combined likelihood estimate. Vertical red lines mark the "true" parameter values and the green shadings mark the prior for each parameter. Note that here we used a uniform prior for the capture probability ($p_d \sim U(0.01, 0.99)$). Above each density plot we also list the median and 95% Credible Interval for the combined single group estimate in blue and the combined likelihood estimate in red.

The coefficient of variation for the posterior is $CV = 1/\sqrt{\alpha + n_p}$, so as $n_p \to 0$ we get $CV \to 1/\sqrt{\alpha}$, thus when $n_p \to 0$ the CV of our estimate approaches that of the prior and as n_p increases, the precision of the estimate improves. The practical implication of this is that the value of CV is directly linked to the number of events (e.g births and deaths) in the period of observation which can be increased by extending the length of the observation period, not necessarily

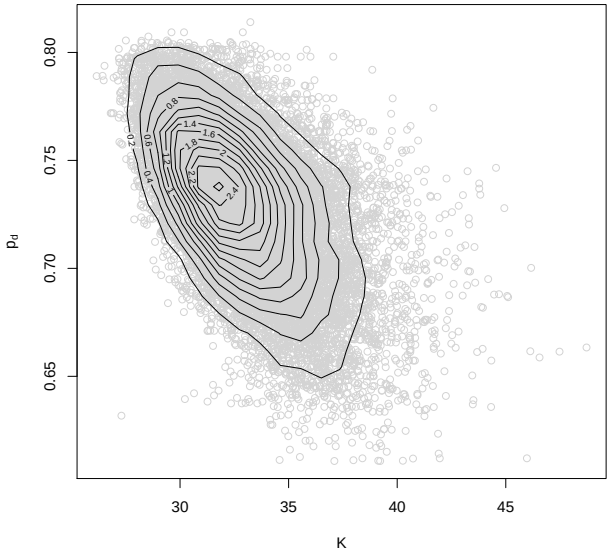


Figure 3.10 Joint density of p_d and K in the estimates with the combined likelihood for 25 groups in Figure 3.9. All data points are plotted as grey circles.

by increasing the observation frequency. Although Equation 3.19 shows that CV does not directly depend on the number of observations, Figure 3.8 and Figure 2 in Catterall et al. (2012) suggest some dependence. This can be explained by the fact that subsequent observations can provide some indirect information about the minimum number of a particular event type. For example in a model with births and deaths only, if the population increases by ΔN this implies that there have been at least that many births in that period (assuming the observations are made without error). If we assume there is only one event type and $n_p \to 0$ then $\beta' \to (t_f - t_0) f(s(t_0))$ which demonstrates that even without any events the initial state $s(t_0)$ contains some information about the parameter of interest p.

Figure 3.8 does suggest that a significant improvement of the estimates of p_d occurs when the observation frequency is increased. This can this be understood by analysing the posterior distribution for p_d and should also apply to the other observation parameters included in Chapter 4. For detection probability p_d we use the following binomial observation model

$$P(D(t)|S(t), p_d) = \prod_{i=1}^{N} \left(\binom{S(t_i)}{D(t_i)} p_d^{D(t_i)} (1 - p_d)^{S(t_i) - D(t_i)} \right) \tag{3.21}$$

where $D(t)$ is the observed state $S(t)$ is the true (inferred) state and N is the number of observations. Using a beta prior for p_d, i.e. $p_d \sim beta(\alpha, \beta)$ the marginal posterior for p_d is

$$P(p_d|D(t), S(t)) = beta(\alpha', \beta') \tag{3.22}$$

where $\alpha' = \alpha + \sum_{i=1}^{N} D(t_i)$ and $\beta' = \beta + \sum_{i=1}^{N} (S(t_i) - D(t_i))$. Equation 3.22

implies that the estimate of p_d becomes less prior dependent as the number of observations N increases and trivially, when we have no observations ($N = 0$), the prior is returned. Note that the argument made in this paragraph for the observation parameters with beta priors also applies if the prior is a uniform distribution as it is just a special case of the beta distribution ($\alpha = \beta = 1$). From the definition of the mean and variance of a beta distribution we can also derive the coefficient of variation for the posterior

$$CV = \sqrt{\frac{\beta'}{(\alpha' + \beta' + 1)(\alpha')}} \tag{3.23}$$

rewriting and assuming that α' and β' are large we get

$$CV \sim \sqrt{\frac{1}{(\alpha'/\beta' + 1)(\alpha')}} \sim \sqrt{\frac{1}{(\alpha')}} \tag{3.24}$$

because $\alpha'\ \beta' \sim O(1)$. And thus since $\alpha' \sim O(N)$ this shows that the coefficient of variation of the observation parameter (i.e. p_d) decays with the number of observations N which could explain why we see such a dramatic improvement in the estimate for p_d in Figure 3.8.

Note that including data from multiple groups appears to affect the process parameters as well as the observation parameters (Figure 3.9). This can be understood because we effectively extend the observation period (sum of the observation periods for all groups) as well as the number of observations (sum of the number of observations made on each group).

3.3.7 Hierarchical prior for the carrying capacity

The assumption that the demographic parameters are the same across groups within a population is obviously a simplification. In particular the carrying capacity of different groups is likely to vary. This is certainly evident in the case of badger social groups which occupy territories of different sizes and quality (e.g. some territories may have better access to food sources and could therefore sustain larger populations). We simulated some data in which the carrying capacity varied from group to group. As to be expected, the resulting estimates using the combined likelihood with these data are very disappointing (Figure 3.11). Apart from the death rate the estimates do not cover the "true" parameter values used in the simulations. For this reason we have implemented a hierarchical prior for the carrying capacity K. In this modification we generate independent estimates of the carrying capacity K_i for each group i from the following posterior distribution conditional on all other parameters as well as the data and the hidden events

$$P(K_i, \mu_K, \sigma_K | D, H, \theta_{-K_i,\mu_K,\sigma_K}) \propto \prod_i L(K_i) P(K_i|\mu_K, \sigma_K) P(\mu_K) P(\sigma_K) \tag{3.25}$$

where the marginal likelihood $L(K_i) = P(D, H|\theta_M)$ as before but now $\theta_M = \{r_b, \mu, K_i\}$. The conditional distribution for K_i is hierarchical, which means that instead of its parameters μ_K and σ_K being fixed they vary and have their own prior distributions.

For the conditional distribution of K_i we use a gamma distribution

$$K_i \sim \Gamma(\mu_K, \sigma_K) \tag{3.26}$$

and we use Gamma priors for μ_K and σ_K:

$$\mu_K \sim \Gamma(\mu_{\mu_K}, \sigma_{\mu_K}) \tag{3.27}$$

$$\sigma_K \sim \Gamma(\mu_{\sigma_K}, \sigma_{\sigma_K}) \tag{3.28}$$

Figure 3.12 shows that estimation of the group carrying capacities K_i performs well. Note that there still seems to be a tendency to overestimate the group carrying capacity K_i but at least now this is bounded below by the maximum observed population in each particular group i and not by the maximum observed population in all groups as in the non-hierarchical version on the inference method.

When using a hierarchical prior for K the other parameters (r_b,μ and p_d) are estimated better than in the case where we use a non-hierarchical inference scheme for K (compare Figures 3.13 and 3.11). The combined likelihood (red curves in plots) for each of the parameters now includes the "true" value. The estimate of p_d based on the combined likelihood has improved from the case where we used the non-hierarchical inference scheme for K but although now it contains the "true" value in its high tail it still seems fairly low. In section 3.3.5 we observed a similar behaviour and we suggested that there is a link between the estimates K and p_d.

3.4 Application to Woodchester Park data

The Woodchester Park study has been conducted by FERA (and it's precursors) since 1975 and a consistent procedure comprising four annual capture campaigns has been established since the early 1980s. Each captured animal is marked with a tattoo and every capture event is included in the database. Apart from general observations regarding the animals age and general state of health, a varying number of samples are taken taken for analysis to establish the disease status (see Chapter 4). The version of the database to which we have access contains information on ca. 3000 individual badgers and many of these have been captured on multiple occasions.

In this chapter we are mainly interested in information on the number of animals present in each social group within the study area and we aim to use this to estimate the parameters assuming the simple model for population dynamics as

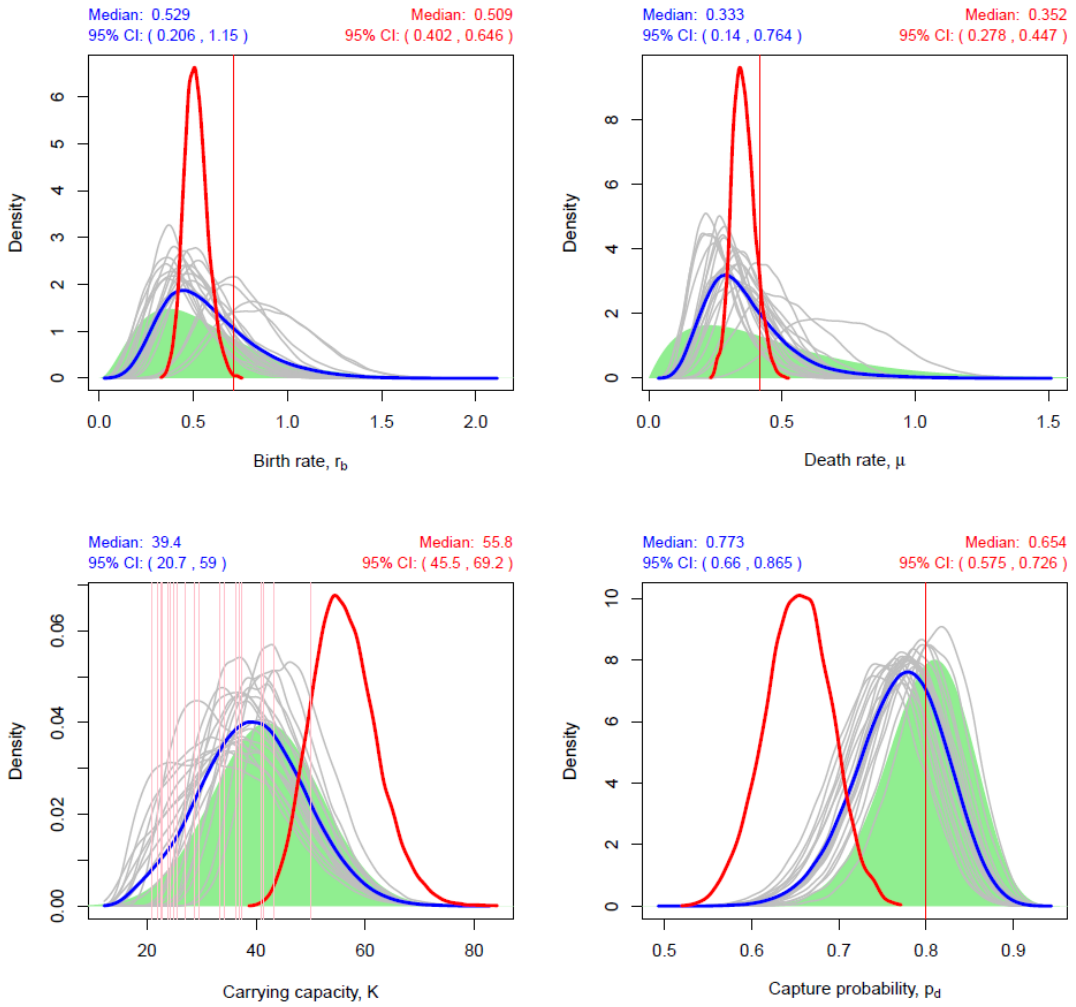


Figure 3.11 Posterior density estimates obtained using the non-hierarchical version of the inference method from simulated data where carrying capacity varies from group to group (as indicated by pink lines subplot for K). Included for comparison, posterior probability densities for 20 simulated data sets (grey) and the combined single group estimate (blue, average of grey curves) of all 20 estimates. A burn-in period of 200000 iterations was applied to all estimates.Vertical red lines mark the "true" parameter values and the green shadings mark the prior for each parameter. Above each density plot we also list the median and 95% Credible Interval for the combined single group estimate in blue and the combined likelihood estimate in red.

outlined before (Section 3.3.1). Although there are four capture campaigns each year, it is not feasible to extract estimates of the population size with a higher than annual frequency from the data base. The main reason for this is that the capture campaigns do not occur at the same time within subsequent years. Our estimates are counts of the number of individuals that were encountered at least once in each year. We make a correction if an individual animal was not a cub in the first year it was caught (counting animals first caught as yearling in the preceding year and those first caught as adults in the preceding two years) and

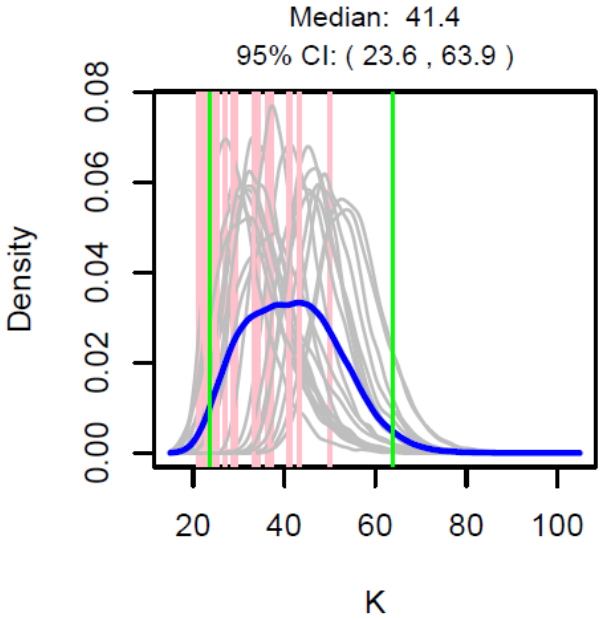


Figure 3.12 Posterior density estimates of the group carrying capacities K_i (in grey) obtained with a combined likelihood for 20 groups and a hierarchical prior for K. The data was the same as in Figure 3.11 (i.e. "true" carrying capacity K_i varies from group to group, pink lines). Note that the blue curve is the average density of K_i and not the combined single group estimate of K that is reported in e.g. Figure 3.13. and the green lines mark the 95% CI for the average density of K_i and the median and 95% Credible Interval of this average density are listed at the top.

also if an individual is not caught for one or more years (we correct the count in the intervening years). Note that this method of establishing the population size from the database takes into account some of the individual based information collected. However, we can not use such information directly as our model is not individual based and we assume the data collected can be used to derive a population size at one particular point in time (end of each year). This is similar to the assumptions made by Catterall et al. (2012) when dealing with plant atlas data. Using these data we are able to extract population histories spanning the period from 1982 to 2008 for about 20 groups. An example of the data that we obtain for a group is given in Figure 3.1(b). Note that using annual "observations" from the database means that the detection probability we use in the observation model is the annual probability of being observed at least once in a particular year and not the probability of being caught in a single campaign. If there are four field campaigns in each year and the probability of being observed during a particular campaign is p_c then (assuming independence) the annual detection probability $p_d = 1 - (1 - p_c)^4$.

We run the inference on the Woodchester Park data with priors that have their mean set at the values that have been used for modelling in the past (see Table 3.1 and Section 3.5). The variance for the priors was chosen to reflect a degree of uncertainty without allowing for estimates that would be unreasonable

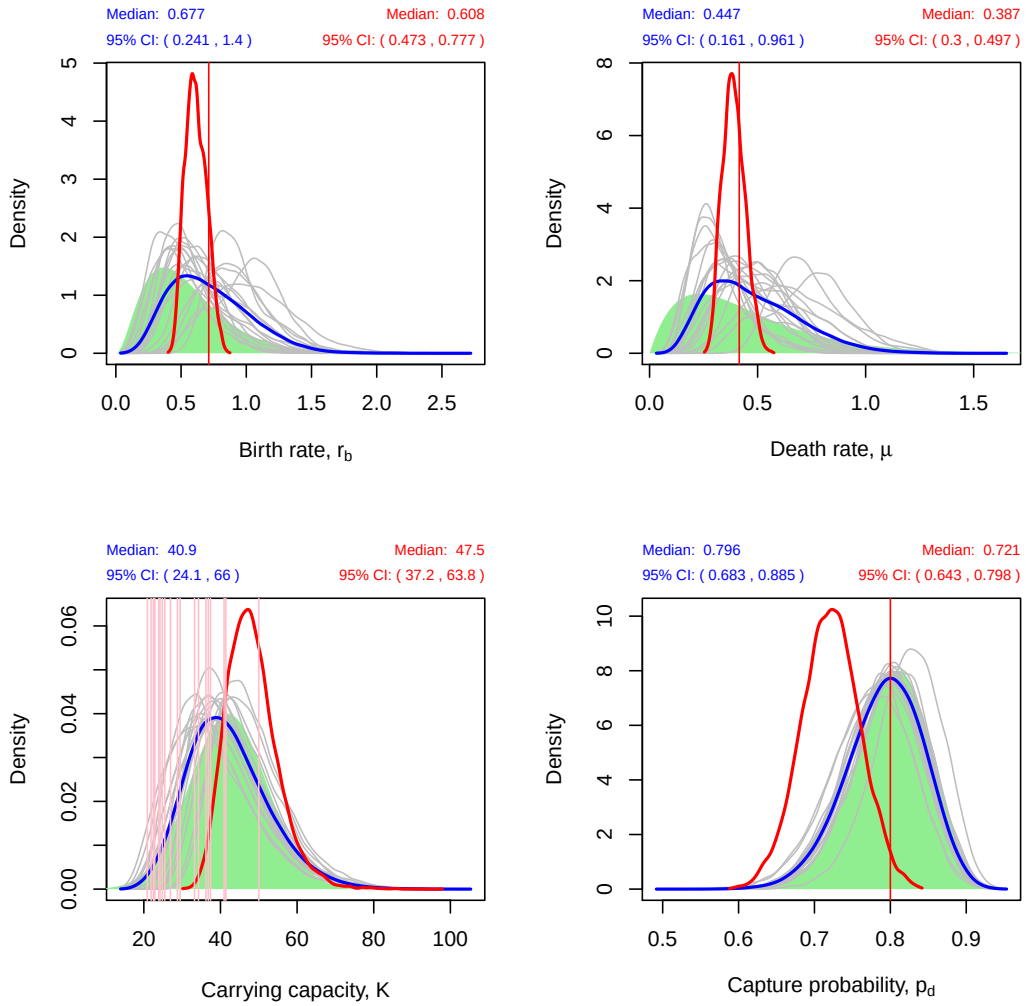


Figure 3.13 Posterior density estimates for all parameters obtained using a hierarchical prior for K from the simulated data with varying carrying capacity (pink lines in plot for K). The red curve is the estimate based on the combined likelihood. Included for comparison, posterior probability densities for 20 simulated data sets (grey) and the combined single group estimate (blue, average of grey curves) of all 20 estimates. A burn-in period of 100000 iterations was applied. Vertical red lines mark the "true" parameter values and the green shadings mark the prior for each parameter Above each density plot we also list the median and 95% Credible Interval for the combined single group estimate in blue and the combined likelihood estimate in red.

in terms of their ecological interpretation. The results are presented in Figure 3.14, Table 3.1 and Appendix B. The combined single-group estimates all appear to be strongly dependent on the prior distributions (green shading in Figure 3.14). This is particularly the case for the birth rate r_b and to a lesser degree for the carrying capacity K and the death rate μ. The estimates based on the combined likelihood (red lines in Figure 3.14) which uses the population histories for 16 groups are more focused for r_b and μ and are considerably higher than the

Parameter	Prior	Posterior estimate
r_b	$\Gamma(\mu_{r_b} = 0.4, \sigma^2_{r_b} = 0.01)$	$0.58(0.45, 0.74)$
μ	$\Gamma(\mu_{\mu} = 0.2, \sigma^2_{\mu} = 0.01)$	$0.40(0.31, 0.51)$
μ_K	$\Gamma(\mu_{\mu_K} = 42, \sigma^2_{\mu_K} = 100)$	$59(45, 78)$
σ_K	$\Gamma(\mu_{\sigma_K} = 10, \sigma^2_{\sigma_K} = 1)$	$10(8.3, 12)$
p_d	$\beta(\mu_{p_d} = 0.8, \sigma^2_{p_d} = 0.0025)$	$0.77(0.66, 0.86)$

Table 3.1 Summary of parameter priors and posterior density estimates for the results on the Woodchester Park data (Figure 3.14). We use the mean and standard deviation to define the priors, and the shape and scale of the gamma and beta priors are re-parameterised from these. The parameter values used in the literature for population dynamics in badgers are set as the prior mean for r_b, μ and the mean of K used in the hierarchical prior. The prior mean for p_d is set to the annual detection probability thought to be achieved in Woodchester Park. The posterior estimates for each parameter are summarised with the median followed in brackets by the 95% Credible Interval.

current thinking about the model parameters as reflected in the priors.

Estimation of the detection probability proved infeasible with the Woodchester Park data. Using "vague" priors for p_d did not result in the estimates converging and the estimates with tight beta priors resulted in a posterior estimate of p_d that was very close to the prior (Figure 3.14). It is hard to estimate this parameter (Delahay et al., 2000) but the consensus in the Woodchester Park team is that every year around 80% of the animals present are caught (pers. com. Gavin Wilson). It is not clear how this consensus estimate is constrained. Therefore we include some variance in the prior for the detection probability to reflect the uncertainty about this parameter.

3.5 Discussion

Our tests on simulated data show that our inference approach performs well on complete data where all events are observed as and when they happen (Figures 3.4 and 3.5). Testing on simulated data with annual observations shows that the method also works. In Woodchester Park we have information from multiple groups which have been observed simultaneously. We can make use of this by (a) combining the posterior estimates of the parameters from different groups by averaging their densities or (b) by combining the likelihoods of different groups which results in a single posterior estimate of the parameters. Combining the single-group estimates results in average parameter estimates which appear to be very closely dependent on the prior distributions (e.g. Figure 3.7 and 3.8). More focused and less prior dependent estimates are obtained by including data from multiple groups (replicates) and multiplying their likelihoods as illustrated by Gibson et al. (2006) (Figure 3.9).

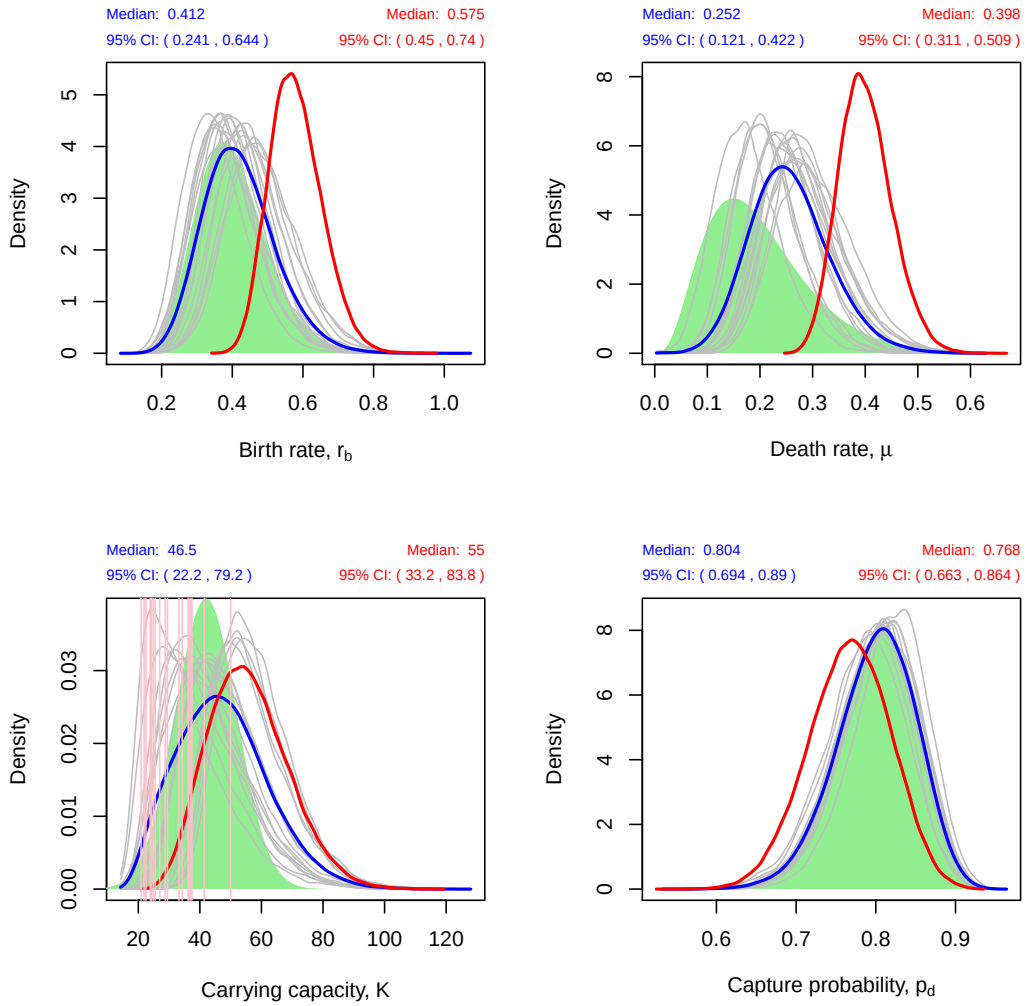


Figure 3.14 Posterior probability density estimates for each of the demographic parameters derived using a combined likelihood for 16 groups (red curve) in the Woodchester Park database (see Appendix B for the full results). Included for comparison are posterior probability density estimates derived using the likelihood for 16 single groups (grey) and the combined single-group estimate (blue, average based of these 16 estimates). A burn in of 500000 iterations was applied. The priors used (green shadings) are centred on parameter values commonly used in the modelling literature and the value of the detection probability assumed to be achieved by the team in Woodchester Park. Above each density plot we also list the median and 95% Credible Interval for the combined single group estimate in blue and the combined likelihood estimate in red.

When combining likelihoods of multiple groups we initially assume that all parameters are identical across groups. In our simple model for population dynamics this assumption may be acceptable for the birth and death rates but it is harder to defend with respect to the carrying capacity K. The carrying capacity of small groups of badgers or other wildlife species with a territorial social structure is likely to vary with territory size and quality (e.g. some territories may have

better food resources than others). We allow for variation of parameters between groups in the combined likelihood estimate by implementing hierarchical priors. In principle this could be done for all parameters but here only implemented this for the carrying capacity K, and this results in very good estimates when we apply the inference method to simulated data with varying carrying capacity across groups (Figure 3.13).

Making the observation model and its parameters an integral part of the inference method in theory also allows the observation parameters to be inferred (Marion et al., 2012). This has been demonstrated on discrete time models (e.g. (Clark and Bjørnstad, 2004; Pagel and Schurr, 2012)) but to our knowledge has not been done successfully on continuous time state space models. In most published cases the observation parameters are fixed either explicitly (e.g. (McInerny and Purves, 2011) or implicitly by not including an observation model (e.g. (Catterall et al., 2012)). We demonstrate in Figure 3.8 that it is possible to estimate the detection probability p_d for our particular implementation as long as we have data that is sampled at a high enough frequency. In practice and in particular in our application to the Woodchester Park data we generally do not have data that is observed with sufficient frequency to do this. Although we may not have enough information to obtain an estimate of the detection probability that differs significantly from the prior distribution, allowing it to be estimated does seem a fairer reflection of what we actually know about the observation process and also seems to result in better mixing of the Markov chain.

Inferring the unobserved state space history is an integral part of our approach and in each iteration we infer a large number of such state space histories that are consistent with the data taking into account the observation model (Figure 3.3). This information can be used to infer the population size at any point in time. A similar method was used by Catterall et al. (2012) to infer the historic distributions of Giant Hogweed in the UK.

Despite these generic methodological developments, our focus here has been on their application to population dynamics processes in wildlife populations and the data from Woodchester Park. In section 3.4 we note that we do not make full use of the individual level data collected at Woodchester Park since our model is not individual based. However, here we focused on population level models since in many wildlife studies reliable individual based observations may not be available. Nonetheless extending our methods to account for individual based information would be an interesting extension of the work presented here. Capture mark recapture models make more use of the individual based information but these tend to be in discrete time and as Buckland et al. (2007) comment continuous time models offer a more realistic representation of demographic processes than their deterministic counterparts.

It is clear that the model we use is somewhat simplistic compared to most of the published models for badger population and disease dynamics (Anderson and Trewhella, 1985; White and Harris, 1995a; Smith et al., 1995) which tend to include gender and age structure, some form of seasonality and some interaction between spatially segregated groups. Most of these models are mass-action models although others are individual based (e.g. Shirley et al. (2003)). The absence of an age and gender structure in our model means that all the rates are *per capita*

and the absence of seasonality means that all the rates are annual and constant. It possible to apply our approach to more complicated models for population dynamics. However, in Chapter 4 we also include a disease component which adds a significant level of complication, but as we demonstrate our inference framework can be extended to this more complex problem. By implementing a hierarchical prior for the carrying capacity K we allowed K to vary from group to group as that is a reflection of the variation of the size and quality of their habitat. Our current assumption is that the other parameters (birth and death rates) are identical from group to group as they are likely to be specific for a species and less sensitive to the local environment. However, it is likely that all of the parameters vary from year to year (e.g. as a response to climatic variation). Such temporal variation is currently not included in our implemented approach and if present may affect our estimates.

With the analysis presented here, we now have obtained estimates for the annual per capita birth rate, death rate and carrying capacity for a simple population dynamics model based on the long term observations of badgers in Woodchester Park 3.4. Making a comparison between our results and the rates used by different authors is not trivial because the modelling approaches differ between authors (an overview of published models is given in Appendix A). In our comparison we focus on the birth and death rates (r_b and μ) and do not include a comparison of the carrying capacity because it is a fairly abstract quantity that can not be measured directly. In our model definition the carrying capacity K is the largest possible group population size and affects only the birth rate. Obviously this depends on the scale at which we model and because we allow the carrying capacity to vary from group to group a comparison does not seem particularly useful. Early models tend to be deterministic and in continuous time (e.g. Anderson and Trewhella (1985)) while later ones are stochastic models in discrete time (e.g. White and Harris (1995a) and Smith et al. (1997)). In terms of model structure there are differences too: in our model density dependence is included only in the birth term through the carrying capacity K, whereas some authors have included various forms of density dependence in the death term as well (e.g. Anderson and Trewhella (1985), White and Harris (1995a) and Smith et al. (1997)). The structure of our model is closest to the version of the Anderson and Trewhella (1985) without age and gender structure.

Previous parameter estimates were mostly based on reviews and compilations of existing observations with respect to general badger ecology (Anderson and Trewhella, 1985; White and Harris, 1995a; Smith et al., 1997) and did not involve estimates of parameter uncertainty other than the ranges quoted in the literature. Swinton et al. (1997) used a model identical to Anderson and Trewhella (1985) and focused only on the Woodchester Park observations and we have used values close to their estimates as the prior mean for the birth and death rate. Previous estimates of birth rates were based on general observations of badger ecology (e.g. fecundity, sex ratio and age structure etc.) whereas death rates were based observations of life-expectancy which can be translated into mortality. A comparison is given in Table 3.2. In summary, our parameter estimates appear to be on the high side of those used in the past but are not unreasonable considering general knowledge about badger ecology. An obvious advantage of our approach is that

Authors	Birth rate	Death rate
Anderson and Trewhella (1985)	0.6	0.4
Swinton et al. (1997)	0.42	0.26
White and Harris (1995a)	-	0.24 - 0.40
Smith et al. (1997)	0.39	0.26
This study	$0.58(0.45, 0.74)$	$0.40(0.31, 0.51)$

Table 3.2 Comparison of birth and death rates used in previous modelling studies results and our estimates based on the Woodchester Park data (Section 3.4). The parameters used by Swinton et al. (1997) are derived explicitly from the Woodchester Park observations and we used values close to these to set the prior means for the birth and death rate in the inference. No birth rate is included in this table for White and Harris (1995a) because in their model it was not included explicitly. The result quoted from our study are the median of the estimate based on the combined likelihood followed by the 95% Credible Interval in brackets.

our estimates include some estimate of the possible variation in the parameter estimate as documented by the 95% CI in Table 3.2 and the plots for the posterior densities in Figure 3.14.

Having derived estimates for the population dynamics parameters including their associated uncertainty it would make sense to use these in a predictive context. This is beyond the scope of this thesis, but an obvious area where this would be useful is the modelling of how a wildlife population recovers from population reduction (e.g. culling). This is an interesting question in itself but also has important implications in the context of TB in badgers and in the light of the results from the Randomised Badger Culling Trial (ISG, 2007) which proved that culling badgers had the effect of increasing the level of disease in cattle as well as the badger population (Jenkins et al., 2007). This is referred to as the "perturbation effect" and was not predicted by any of the existing models (e.g. (White and Harris, 1995a; Smith et al., 1997)). Recent work by (Prentice, 2012) showed that under specific conditions the perturbation effect can result from very simple models without explicitly changing the parameters at the point of culling. Obviously such predictions would benefit from including the parameter estimates from this study which would allow some measure of uncertainty for the outcomes to be reported as well.

3.6 Conclusions

Here we have extended an existing integrated approach for Bayesian MCMC parameter inference for continuous time discrete state space models (as documented by Marion et al. (2012), Figure 3.2). We implemented the inference method for a simple model of population dynamics in small unstructured wildlife groups affected only by births and deaths and tested it with simulated data. We have

(a) incorporated an observation model where the observation parameters are also inferred, (b) combined the likelihood for multiple groups to decrease the dependence on the prior and (c) developed a hierarchical scheme to allow for variation in parameter values between groups. The method was applied to the data from the long term observations of a population of European badgers in Woodchester Park (SW England) and are able for the first time to present estimates of the birth and death rates with their associated uncertainty. The estimated median and 95% Credible interval are $0.58(0.45, 0.74)$ for the birth rate and $0.40(0.31, 0.51)$ for the death rate.

Chapter 4

Parameterising models for population and disease dynamics

4.1 Introduction

There are many and varied contexts in which understanding disease in wildlife populations is important including: wildlife management and conservation (Daszak et al., 2000; Deem et al., 2001), human public health (many diseases are zoonotic and can affect multiple host species (Claas et al., 1998; Smith et al., 2005; Jones et al., 2008; Kubiak and McLean, 2012)), and agriculture (Donnelly et al., 2006; Judge et al., 2007; Davidson et al., 2009). The problem of Tuberculosis (*Mycobacterium bovis*) in badgers (*Meles meles*) and cattle is a clear example of how disease in wildlife can affect agricultural practice and can have a major economic impact (e.g. DEFRA (2004); ISG (2007)). Several population and disease dynamics models for TB in badgers have been published (e.g. Anderson and Trewhella (1985); Bentil and Murray (1993); White and Harris (1995a); Smith et al. (1995); Barlow (1996); Swinton et al. (1997); Shirley et al. (2003), for a tabular summary see Appendix A) but none of these studies attempted to estimate model parameter values and the uncertainty associated with those estimates by integration of the models with data from observations. The models published so far use fixed parameter values ignoring any uncertainty associated with their estimation. Moreover, these parameter estimates are obtained by treating observations of particular ecological and epidemiological processes in isolation. However, these processes are component parts of a larger interacting system and the integrated approach we adopt here infers parameter values for each subprocess whilst taking account of the modelled system as a whole. We adopt a Bayesian approach to inference which enables prior information about component model parameters to be combined with observations of the system. To achieve this, models of key processes are combined to develop a process-based model of the system and this is then integrated with an observation model and the prior information to derive a combined distribution across all model parameters. This Bayesian posterior distribution represents the information we have about the model parameters given the model structure, the available data and the prior knowledge about the values of the parameters. This approach to inference was presented in Chapter 3

for a generic discrete state-space Markov process and then applied to modelling demographic fluctuations in wildlife populations.

As stated at the beginning of Chapter 3, disease dynamics in small groups of wildlife species can be strongly affected by fluctuations in population size (Anderson and May, 1979; Wilson and Hassell, 1997) and this is particularly important in small populations. Chapter 3 demonstrated that the approach to inference outlined above can be successfully applied to infer estimates of the parameters for a simple model for population dynamics with births and deaths only. In this chapter, this approach is extended further to apply it to a model that also incorporates epidemiological processes including two types of disease transmission as well as the transition from the exposed state (sometimes referred to as the latent state) to the infectious state and all parameters are inferred simultaneously. The introduction of these disease components requires the careful development of a suitable observation model. In this chapter we discuss in some detail how modelling the observation process involves understanding the capture process, as well as the characteristics of two disease tests and how these are used to make the link between the observed and inferred states in the model.

As in Chapter 3 we use simulated data in order to check how well the inference method performs on each of the parameters. We show that the inference method is successful for simulated data from a model with both population and disease dynamics processes in cases where the disease prevalence is not very low. Application to the data for TB in badgers from Woodchester Park is discussed briefly at the end of the chapter.

4.2 Methodology

In this chapter we use the generic inference approach as outlined by Marion et al. (2012) and extend the implementation in Chapter 3 to include both population and disease dynamics processes. In this chapter we apply the same qualative criteria as in Chapter 3 to judge whether an inference run has converged.

4.2.1 A simple model for population and disease dynamics

We apply the modelling approach for continuous time state space models outlined by Cox and Miller (1965) and Gillespie (1977) and described in Chapter 3 to formulate a model that includes (a) population dynamics processes (i.e. births and deaths) and (b) disease processes (i.e. transmission, background transmission and transition from exposed to infectious state). In this approach we define the rate at which each possible type of event e_i occurs as a function of the state $s(t)$ of the system and the set of parameters θ_M that govern the process. We adopt the same model for population size $N(t)$ at time t as in Chapter 3 with the modification that the population N is divided into three disease states. Individuals in this model have either (a) not been exposed to the disease and are classed as susceptible S, or (b) been exposed but are not infectious (yet) E or (c) are infectious I (Figure 4.1). Note that in this context the term exposed has the technical meaning that a pathogen has invaded the host individual but has yet

to reached the point where it is able to use that individual to infect other hosts. A susceptible individual may have been in an environment containing pathogens but so far has resisted them and failed to become 'exposed'. The total population of a group is made up of the sum of the numbers present in each of the disease states (i.e. $N = S + E + I$).

Birth events e_b occur at logistic rate $r(e_b|N(t);\theta) = r_b N(t)(1 - N(t)/K)$ and newborns are added to the susceptible class, where r_b is the birth rate and K is the carrying capacity. Deaths e_d occur at the same rate for all disease classes $r(e_d|N(t);\theta) = \mu N(t)$, where μ is the death rate. We assume a homogeneous disease transmission network where all susceptible individuals at any particular point in time have an equal probability of being infected. Transmission events e_t occur at rate $r(e_t|S(t), I(t)) = \beta S(t)I(t)$ where β is the transmission rate. Background transmissions e_{bt} are included to allow for recurrence of disease after local disease extinction. Events of this type occur at rate $r(e_{bt}|S(t)) = \phi S(t)$ where ϕ is the background transmission rate. The final event type e_{EI} is the change from the exposed state E to the infectious state I and this occurs at rate $r(e_{EI}|E(t) = \tau_{EI}E(t)$ where τ_{EI} is the rate at which exposed individuals E progress to become infectious I. In summary, in this version of the model we have three state-space variables (i.e. $s(t) = \{S(t), E(t), I(t)\}$) and the set of model parameters is $\theta_M = \{r_b, K\, \mu, \beta, \phi, \tau_{EI}\}$. All rates are annual and *per capita* rates and thus no seasonality or structure other than the disease classification are included.

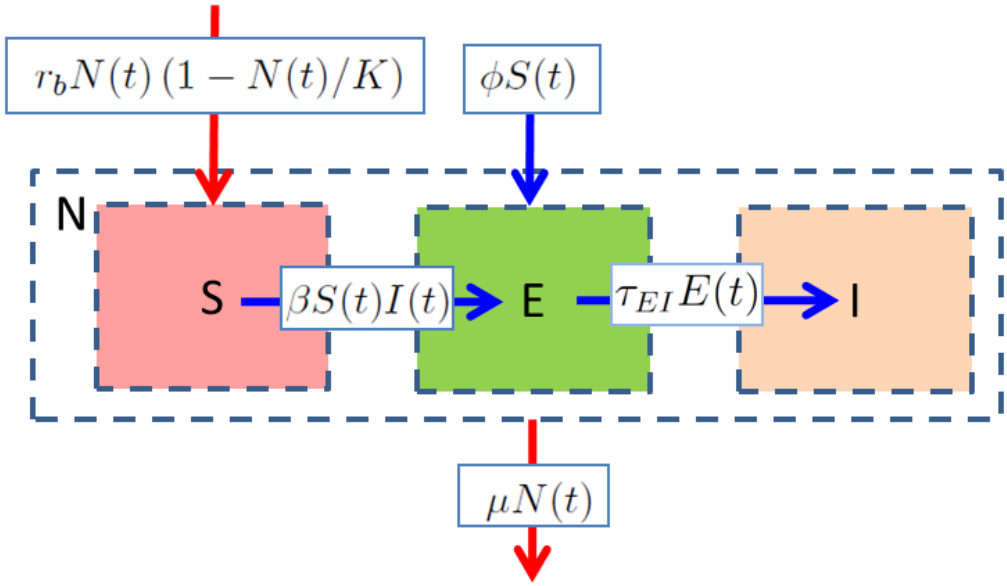


Figure 4.1 Process model including demographic and disease processes, details explained in the text.

4.2.2 Data characteristics

One difference between the application of the inference technique to the model for population dynamics only (Chapter 3) and the model for combined population and disease dynamics is that in the latter case we have a more complicated data structure that has to be captured in the observation model as illustrated in Figure 4.2.

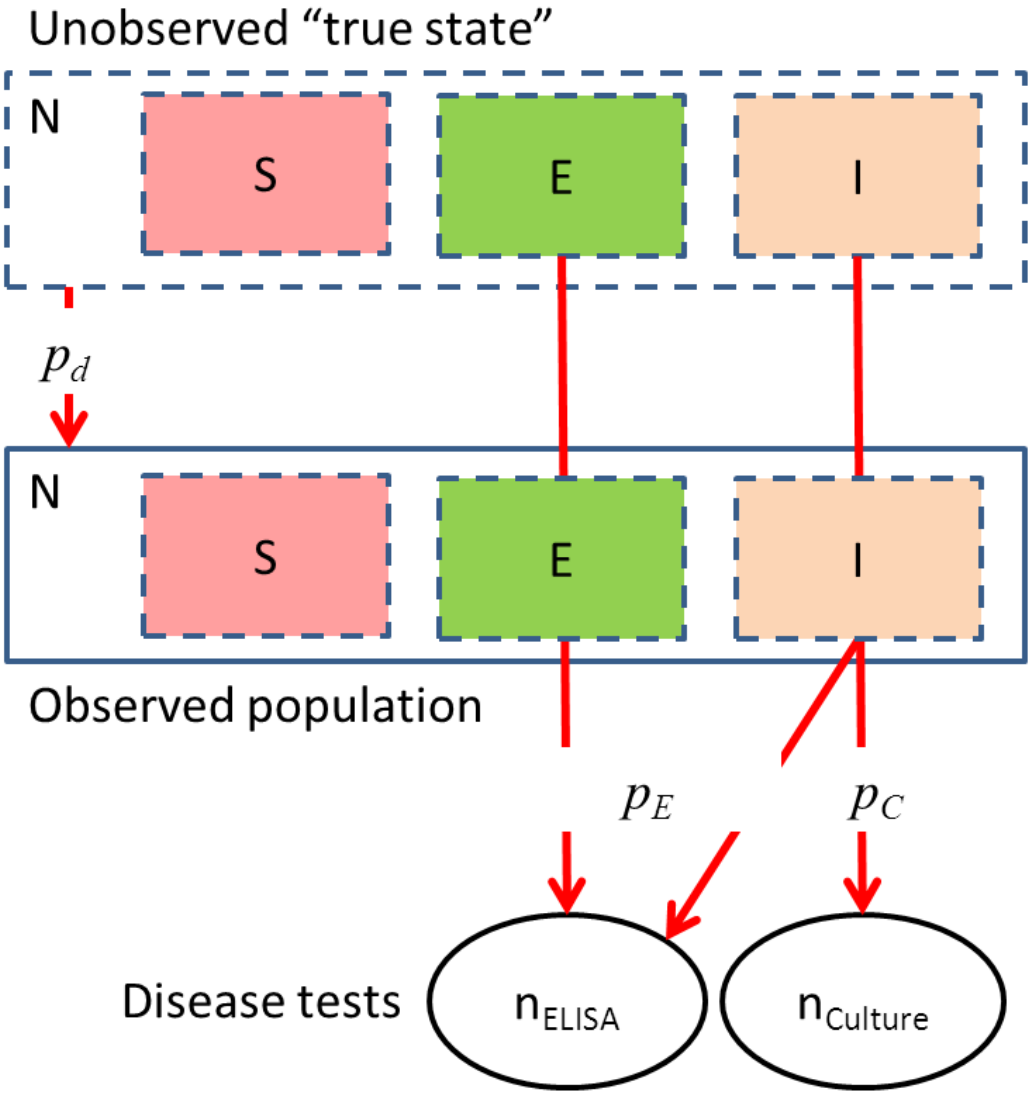


Figure 4.2 Observation model. Observed variables indicated by solid outlines are the population size N and the number of positives in each of the tests (n_{ELISA} and $n_{Culture}$). Variables marked with a dashed outline are not observed directly. Individuals present in the population are observed with detection probability p_d and the numbers of positives in the tests depend on the numbers observed in the disease states E and I and the test sensitivities p_E and p_C. Details are explained in the text.

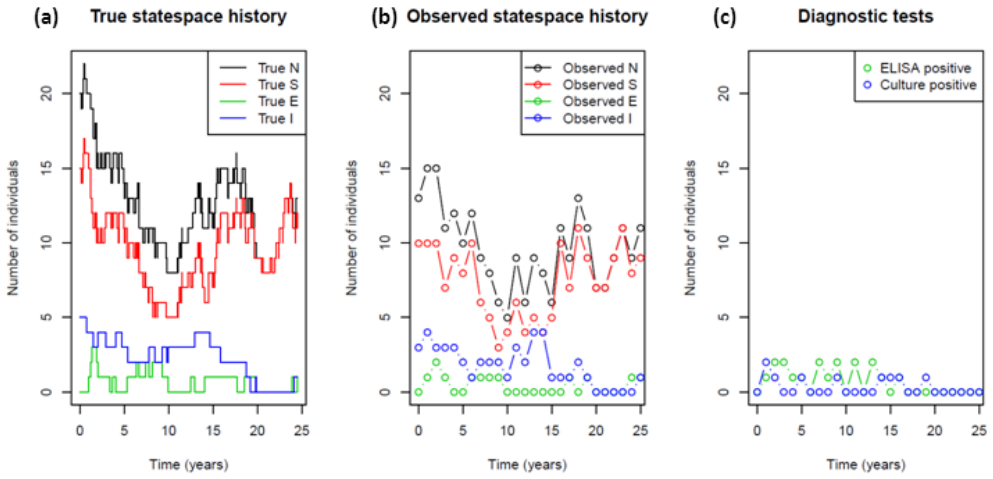


Figure 4.3 Example of data simulated with a model for population and disease dynamics with three disease states (SEI) and an observation model that includes the detection probability p_d and disease test sensitivities p_E and p_C. The "true" state space history is recorded in (a). The observed state space history in (b) shows the state space variables after the application of a binomial detection probability ($p_d = 80\%$). Note that the numbers in each of the disease states (SEI) are those included in the observed population but are not actually observed directly (see text). (c) Shows the number of positives obtained after applying the test sensitivities (in this example we used $p_E = p_C = 30\%$) to the "observed" state space history shown in (b).

Figure 4.3 shows a simulated data set and illustrates some of the issues we need to deal with when applying our inference approach to a real data set such as Woodchester Park. In the case of simulated data we have full knowledge of the state space history and we can observe the state space variables continuously and without error (Figure 4.3a). In real scenarios, ecological and epidemiological systems are observed at discrete points in time (not necessarily evenly spaces) and the observations generally include some error. Errors in observations of the population size may depend on the ecological system or species of interest and also on sampling effort. Figure 4.3b shows an example from the same simulation as Figure 4.3a but with individuals from the population as a whole and the different disease classes observed with an 80% binomial detection probability (see Figure 4.2, and more detail on the observation model in Section 4.2.3). However, the number of captured animals in each disease category is not measured directly because animals are classified as affected by disease on the basis of disease tests. In Figure 4.3c we therefore show an example of simulated disease test results. We incorporate two types of test and apply each once to all of the individuals captured at any given point in time regardless of their actual disease state (Figure 4.3c). Note that since our model is a population level description of the system we can't ascertain whether or not the same individual is caught and tested in different sampling events (i.e. at different points in time). The ELISA is a serology test

that detects antibodies in blood samples and we assume that both exposed and infectious individuals can test positive with sensitivity p_E. The culture test can be applied to a variety of substances sampled from the animals (sputum, urine, etc.) and is thought to show if an individual is infectious with sensitivity p_C (Figure 4.2). In the example in Figure 4.3c we use an illustrative value of 30% for both p_E and p_C. In order to keep the observation model simple we assume that there are no false positives and thus test specificities are ignored (or implicitly set to 1). This is in line with the consensus within the Woodchester Park team (pers. comm. Dez Delahay).

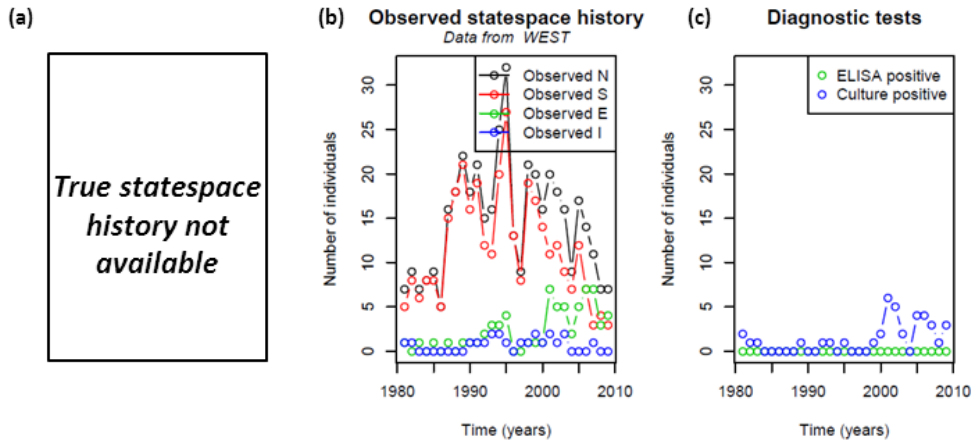


Figure 4.4 Data extracted from the Woodchester Park database for the social group referred to as West. The true state-space history (a) is not recorded. The observed state space history (b) is an annual count of the size of each disease class. The test results (c) reflect the number of animals testing positive at least once in a particular year to the ELISA and Culture tests. Note that although the information in the database is individual based we can not use this information because our model isn't individual based.

In the case of a real data set (an extract for a single sett in Woodchester Park is shown in Figure 4.4) we do not have full knowledge of the state space history (Figure 4.4a). The data is individual based and comprises capture records of individual animals with test results as well as classification into the different disease categories S,E and I. The population is observed in four campaigns every year but because these campaigns are not at fixed times in the year it is infeasible to extract summary data at a higher than annual frequency for the number of animals in each of the disease states (Figure 4.4b)and the number of positives in the tests(Figure 4.4c). An animal is counted if it is caught at least once in a particular year with a small correction to account for individuals that were not first caught as a cub or with a gap in their capture record (see Chapter 3). These counts are sometimes referred to as the *minimum number alive*. No such corrections are applied to the disease test counts. Figures 4.4b and c clearly show that the levels of disease in Woodchester Park are low, raising the issue

that such data may not contain enough information to allow the parameters to be estimated, as discussed in Section 3.3.6.

4.2.3 A specific observation model

The observation model for the population is straight forward (Figure 4.2) and is identical to the observation model used in Chapter 3. As there is only a single observation process (capturing) we can simply compare the observed population $\hat{N}$ with the inferred "true" population N through a binomial probability distribution, i.e. $\hat{N} \sim Bin(N, p_d)$, where p_d is the probability of capturing an individual.

The observation model for the disease component is more complicated because we have to take into account a combination of two observation processes in conjunction (capturing and testing for disease). As discussed in Section 4.2.2 we observe the number of positives in the ELISA test n_E and the culture test n_C. The observation model must link these test results to the inferred state space history.

The observation model for the culture test is the probability of having n_C positives given the "true" number of infectious individuals I and the observed population size $\hat{N}$:

$$P(\hat{n}_C|I, \hat{N}) = \sum_{\hat{I}=n_C}^{I} P(\hat{n}_C|\hat{I})P(\hat{I}|I, \hat{N}) \tag{4.1}$$

The first term in this sum reflects the disease testing process and is a straightforward binomial probability, i.e. $P(\hat{n}_C|\hat{I}) = Bin(\hat{I}, p_{C,ann})$. The second probability is the probability of infected individuals $\hat{I}$ being included in the observed population $\hat{N}$ given the true number of I and the observed population $\hat{N}$. This depends on the probability that an individual in the true population N is infected I/N as well as the probability of capturing that individual p_d, so that the probability of detecting an infected individual is $p_d I/N$ and therefore $P(\hat{I}|I, \hat{N}) = Bin(\hat{N}, p_d I/N)$. Because we assume the test specificity equals 1 (no false positives) the number of infectious $\hat{I}$ included in the observed population $\hat{N}$ can be anywhere between n_C and the true number of infectious individuals I.

The observation model for the number of ELISA tests takes a similar shape to the model for the culture test, but because we assume that both exposed and infected individuals can test positive in the ELISA with equal probability, we get

$$P(\hat{n}_E|E + I, \hat{N}) = \sum_{\hat{E}+\hat{I}=n_E}^{E+I} P(\hat{n}_E|\hat{E} + \hat{I})P(\hat{E} + \hat{I}|E + I, \hat{N}) \tag{4.2}$$

where $P(\hat{n}_E|\hat{E} + \hat{I}) = Bin(\hat{E} + \hat{I}, p_E)$ and $P(\hat{E} + \hat{I}|E + I, \hat{N}) = Bin(\hat{N}, p_d(E + I)/N)$.

4.3 Results

4.3.1 Estimation of disease parameters on data from single groups

We begin the description of the results with an illustration using simulated data with high disease prevalence (up to ca. 50%) (Table 4.1) compared to Woodchester Park where the group prevalence I/N (based on reported disease classifications) is generally quite low and rarely exceeds 20% resulting in low numbers of positives in the disease tests. We first run the inference to estimate the disease parameters only on annually observed simulation data for a single population (Figure 4.5). In this case we assume that the birth and death rate as well as the detection probability and test sensitivities are known (i.e. they are fixed at their "true" values and not inferred). Figure 4.5 illustrates that when disease is included we have to infer the history of three state space variables (S, E and I). In general there appears to be a good visual match between the "true" history (red) and the inferred history (grey) of each of the disease states (Figure 4.5). We simulated these data with an 80% detection probability and thus at the observation points the observed population size must be below or at least equal to the inferred population (not shown here but an example is given in Figure 3.3). The annually observed history of the disease states (blue crosses in Figure 4.5) is not always below or equal to the inferred history of the disease states but this apparent inconsistency can be explained by the fact that disease states are not directly observed but are based on interpretation of disease tests (as explained in Section 4.2.3). The number of positive culture tests observed over time (green dots) is consistent with the inferred history of the number of infectious individuals (bottom plot in Figure 4.5c). A similar relationship holds for the number of positive ELISA tests and the sum of the number of exposed and infectious individual as required by the observation model.

Although r_b and μ are fixed, we do infer birth and death events as part of the state space history (Figures 4.5 and 4.6). The estimates of the number of each event type within the inferred state space histories agrees well with the "true" number that were present in the complete simulated data set (Figure 4.6). The estimates for the disease parameters convincingly cover the true parameter values (red lines in Figure 4.7) . The transmission rate β, and to a lesser extent the latency rate τ_{EI}, seem reasonably independent of the priors used (Figure 4.7). In the case of τ_{EI} the assessment of prior dependency is difficult using only the specific results in Figure 4.7 because the chosen prior was centered on the "true" value. In the case of the background transmission ϕ the estimate appears strongly dependent on the prior. The reason for this is that this particular simulation run did not include many background infection events (red line in Figure 4.6d) and thus we should not expect to infer this parameter with great precision (an explanation for this was given in Section 3.3.6). This explanation implies that our estimation approach may struggle generally when the number of events of a particular type is small, reducing the chance of successfully applying our approach to the Woodchester Park data in which disease prevalence is generally quite low (e.g. Figure 4.4). Overall Figure 4.7 shows that our approach works well. Summaries

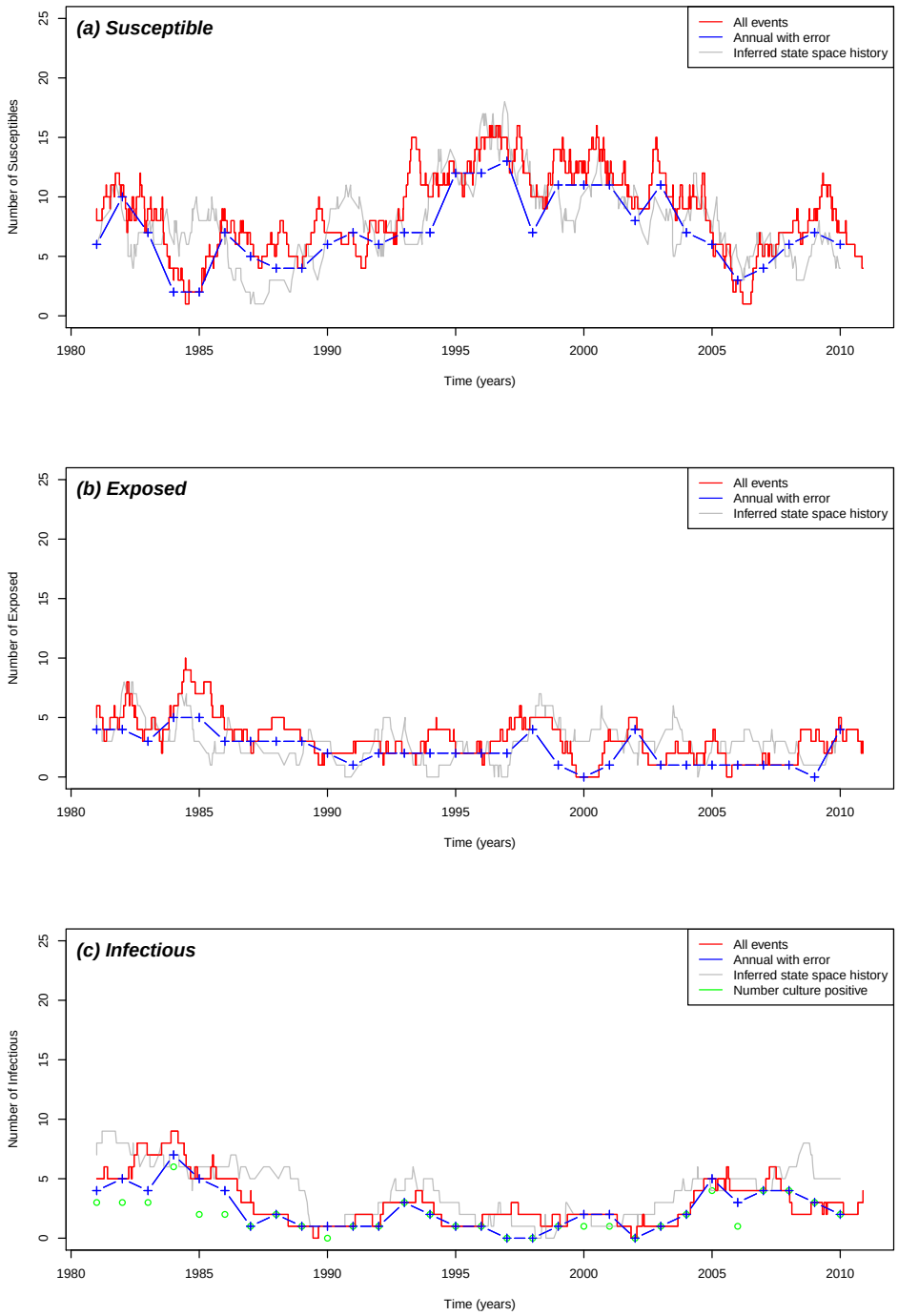


Figure 4.5 Illustration of the relationship in a simulated data set between complete (red), observed (blue) and one example of an inferred history (grey) for each disease state. Note that the number in each disease state is strictly speaking not observed but includes individuals classified as such based on the disease tests. The number of positives in culture tests is given in the lower plot (green circles), number of positives in ELISA is not plotted as it should be compared with both the Exposed and Infectious counts.

	Low prevalence	High prevalence
Population dynamics:		
Birth rate r_b	0.712	0.65
Death rate μ	0.415	0.45
Carrying capacity K	$\sim U(20, 50)$	$\sim U(40, 70)$
Disease dynamics:		
Transmission rate β	0.02	0.1
Background transmission rate ϕ	0.01	0.05
Latency rate τ_{EI}	0.5	0.41
Observation model:		
Detection probability p_d	0.8	0.8
ELISA sensitivity p_E	0.3	0.8
Culture sensitivity p_C	0.8	0.8

Table 4.1 Parameters used to generate simulated data. The low disease content simulations are meant to visually resemble the Woodchester Park data (e.g. 4.4). The population dynamics parameters are based on early inference results on the WP data (not the final results presented in Chapter 3). In the low disease simulations, disease and observation parameters were chosen to be close to published values (e.g. White and Harris (1995a); Smith et al. (1997)). The high disease simulations use arbitrarily high values for the disease transmission parameters and observation parameters that take into account the effect of four captures in each year.

of the estimates (median, 95% and histograms and density plots) should exclude the data from the transient period (burn-in) before convergence is reached but this has not been done in Figure 4.7 as it was intended to illustrate how well the Markov chain mixed.

4.3.2 Simultaneous estimation of demographic and disease parameters on data from single groups

In Figure 4.8 we show the results of an inference run on annually observed simulated data for a single group, where all the process parameters (i.e. the population dynamics parameters r_b, μ and K as well as the disease parameters β, ϕ and τ_{EI}) are inferred simultaneously, but not the observation parameters (p_d, p_E and p_C) which were fixed at their "true" values. The result is encouraging as the 95% CI of the estimates for each of the parameters includes the "true" parameter value (red lines in Figure 4.8). As in the case where only the population dynamics parameters were inferred, the estimate for the background transmission ϕ (Figure 4.8e) rate seems to reproduce the prior. However, the estimates of the other parameters appear less dependent on the prior (Figure 4.8).

The ultimate goal in this chapter is to estimate all population and disease parameters simultaneously as well as the observation parameters and Figure 4.9 shows the relevant results. The complete output for this run is included as Appendix C which shows that convergence is rapid although it could be argued that

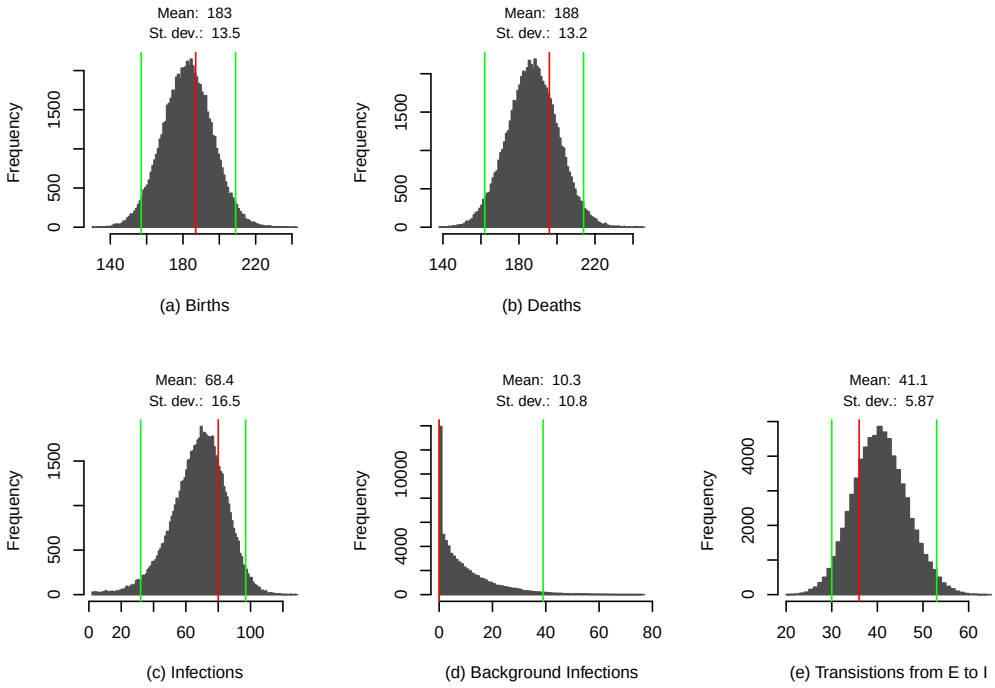


Figure 4.6 Estimates of the number of events of each type obtained using annually observed simulated data for a single group. The values of r_b, μ, p_d, p_E and p_C are not inferred (fixed) but the number of births and deaths are inferred as part of the state space history (e.g. Figure 4.5). The true number of each event type in corresponding complete event sequences is plotted in red, and green lines show the boundaries of the 95% CI. Above each histogram we list the mean and standard deviation of the number of each event type. Note that in the histogram for the transitions from E to I the "true" number coincides with the upper limit of teh 95% CI and the red line overlies the green line.

some of the parameters and event counts only stabilised after about 70000 iterations. In summary this particular example shows the strength of the approach as all parameters apart from the detection probability p_d are well within the 95% CI (Figure 4.9) and proves how much information is contained in fairly limited observations.

In Figure 4.10 we compare the estimates of the disease parameters (β, ϕ and τ_{EI} from the different scenarios presented in this chapter. The estimates for both the background rate ϕ (Figure 4.10b) and the latency rate τ_{EI} (Figure 4.10c) both appear to be very prior dependent. Figure 4.10a shows that we are more successful in estimating the transmission rate β. The best and "most focused" estimate is obtained when inferring the disease parameters only (red curve) (keeping all the other parameters fixed). Obviously this is not possible when using real data because in that case all parameters are not known and have to be inferred. In the scenario where we estimate the population and disease dynamics parameters simultaneously while keeping the observation parameters fixed, the transmission

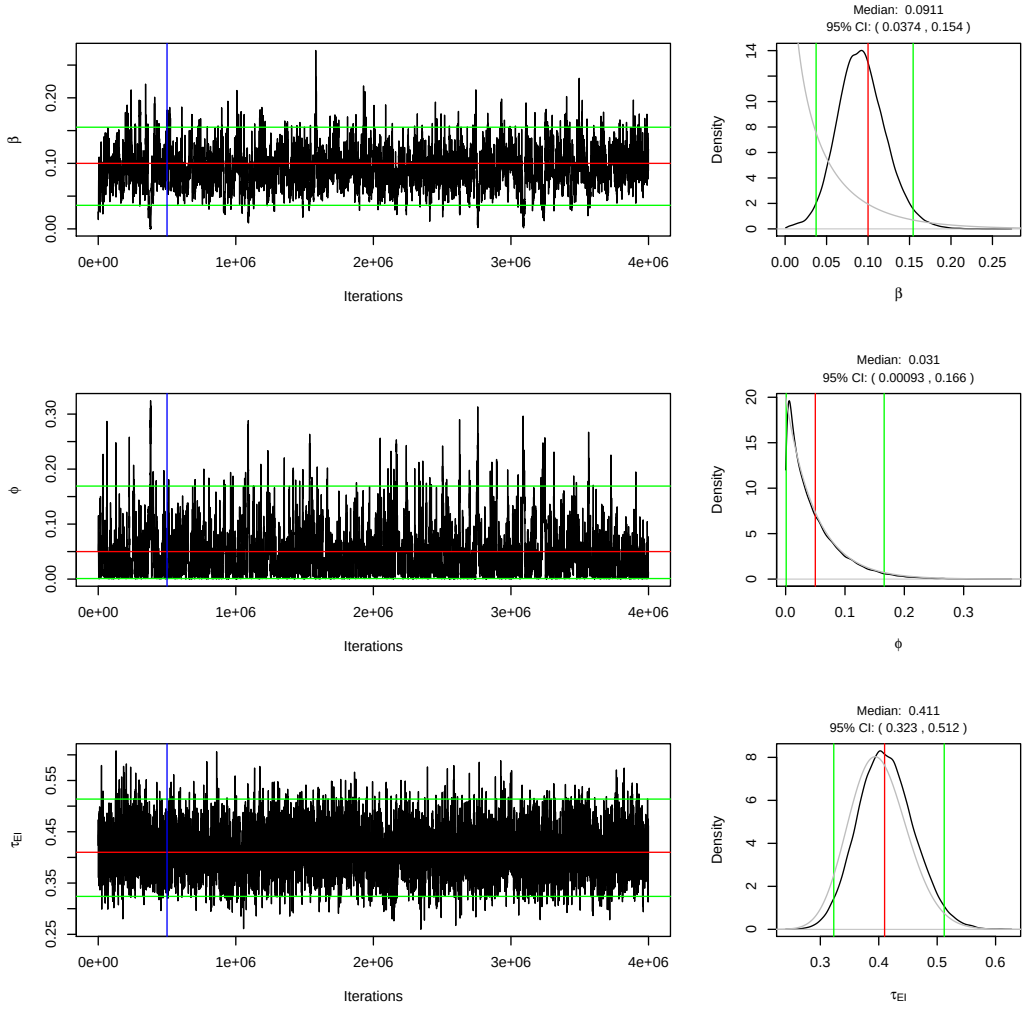


Figure 4.7 Inference results for the disease parameters using annually observed simulated data for a single group obtained with fixed r_b, μ, p_d, p_E and p_C (not inferred). Shown on the left is the trace for the sample and on the right a summary of the density. For reference we have also included the "true" value of each of the parameters used in the simulation (red line), the prior density (grey curve in density plots). The green lines indicate the boundaries of the 95% Credible Interval for each parameter, which is also listed above the density plots together with the median of the estimate. Note that although r_b and μ are fixed, birth and death events are estimated as shown in Figures 4.5 and 4.6. A "burn-in" period of 500000 (blue lines on trace plots) was applied to the density plots.

rate is also estimated well (blue curve in Figure 4.10a). Interestingly the variance of the estimate appears to decrease when we also estimate the observation parameters (green curve in Figure 4.10a) indicating that it is better to allow for the uncertainty associated with the observation processes than to assume exact knowledge about these.

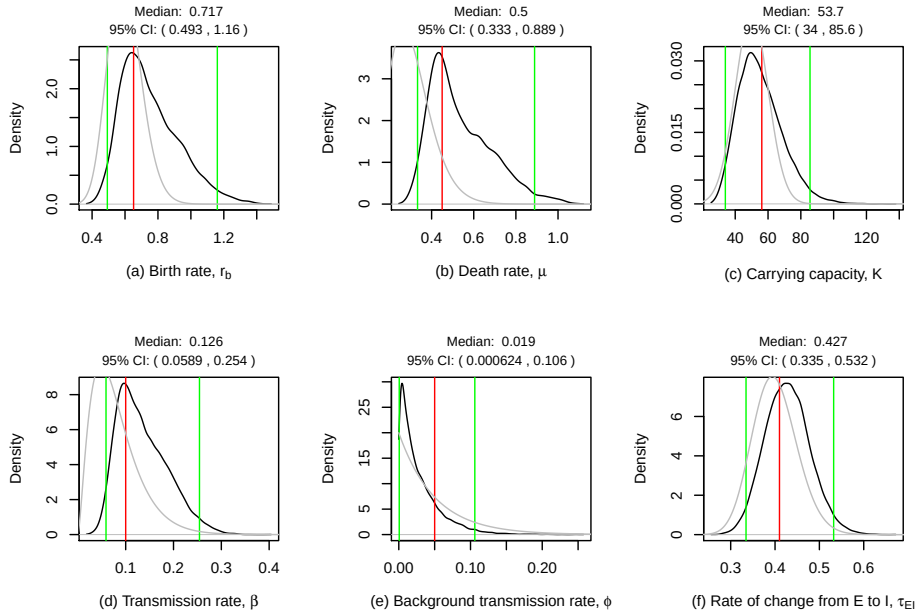


Figure 4.8 Simultaneous inference results for population and disease dynamics parameters on annually observed simulated data for a single group. All process parameters (r_b, μ, K, β, ϕ and τ_{EI}) were estimated simultaneously while observation parameters (p_d, p_E and p_C) were not inferred (i.e. fixed at the "true" values used in the simulation). Included for reference: "true" value of each of the parameters used in the simulation (red line), the prior density (grey curve), boundaries of the 95% Credible Interval (green). Above each density plot we also list the median and standard deviation of the estimate. A burn-in period was not applied.

4.4 Discussion and conclusions

Using simulated data we show that the proposed inference approach (Chapter 3) can successfully be extended to simultaneously estimate all process and observation parameters in a simple model for population and disease dynamics that also includes three observation processes. We implemented this inference approach for a combined population and disease dynamics model and with an observation model that includes both a capture process as well as two disease tests (ELISA and culture tests). Here we demonstrate the feasibility of this approach using simulated data for a single group with high disease prevalence. Despite making several simplifications with respect to real ecological processes, field observations and disease testing practices, our approach represents considerable progress towards simultaneous estimation of population and disease dynamics parameters in the presence of significant observation errors for data from natural systems (e.g. the data on TB in badgers from Woodchester Park).

In Chapter 3 we illustrated, using simulated data, that the best estimates for the population dynamics parameters are obtained when combining the likelihoods for multiple groups as the combined single group estimates tended to be very

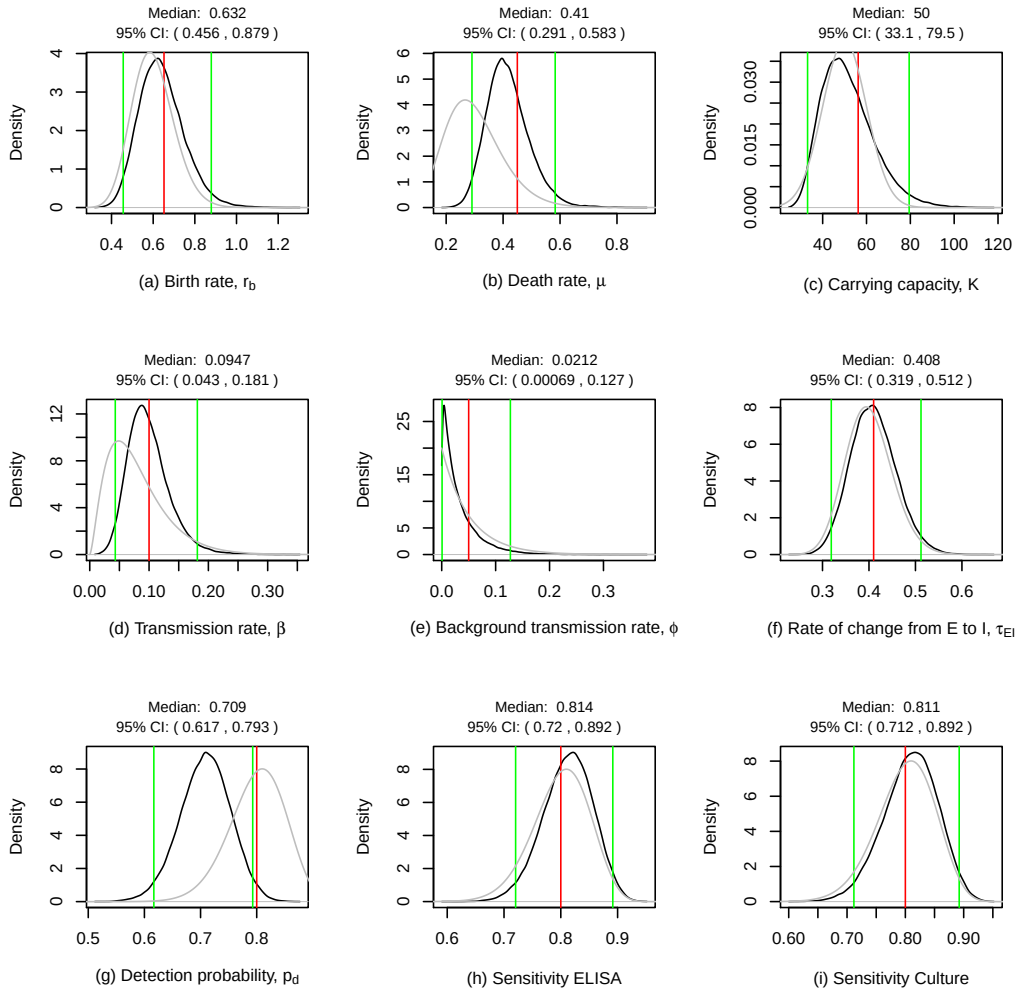


Figure 4.9 Simultaneous inference results for annually observed simulated data for a single group. All process parameters (r_b, μ, K, β, ϕ and τ_{EI}) and observation parameters (p_d, p_E and p_C) were inferred simultaneously. Included for reference: "true" value of each of the parameters used in the simulation (red line), the prior density (grey curve), boundaries of the 95% Credible Interval (green). Above each density plot we also list the median and standard deviation of the estimate. A burn-in period was not applied.

close to the prior and therefore not much weight should be attached to them (Figure 3.12). In Section 4.3 of this chapter we showed for a single simulated dataset with high disease prevalence that we can estimate the posterior density for all the process and observation parameters simultaneously but we have not yet managed to demonstrate that the resulting estimates obtained by combining the likelihood from multiple groups converges as required and whether it leads to the expected improvements. The results so far (not presented in this thesis) look promising but the rate of convergence is very slow and it appears that more tuning of the inference parameters is needed. The rate of convergence depends

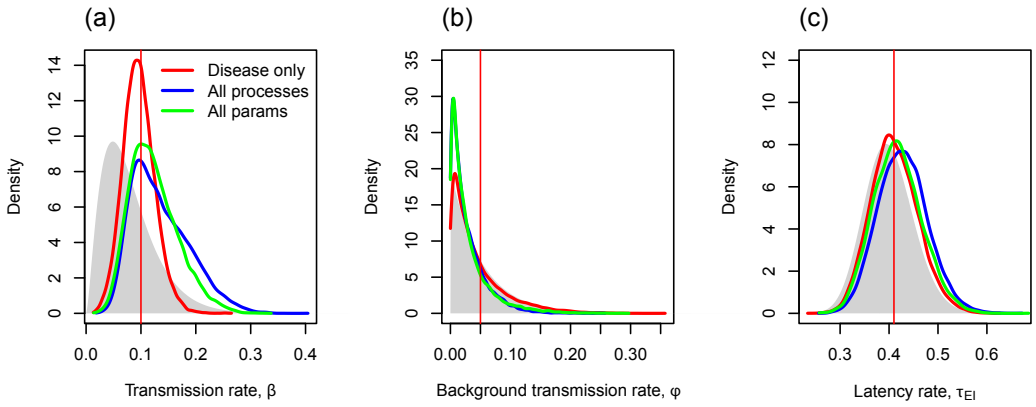


Figure 4.10 Comparison of disease parameter estimates (β, ϕ and τ_{EI}) obtained using 3 different inference scenarios on a simulated data set for a single group: (1) inference of only disease parameters (red) with population dynamics and observation parameters fixed at their "true" values (Figure 4.7), (2) simultaneous inference of population and disease dynamics parameters with observation parameters fixed (blue) (Figure 4.8), and (3) simultaneous inference of all process and observation parameters (green)(Figure 4.9). Included for reference: "true" value of each of the parameters used in the simulation (red line) and the prior density (grey shading). A burn-in of 200000 was applied.

strongly on the number of changes made in each iteration in the inference run. Determining the optimal number of changes is a fairly ad-hoc case of trial and error, although it appears from the work to date that it depends on the number of events in the complete data set (obviously we only know this when working with simulated data), as well as the number of event types and possibly also on the number of observation processes.

The ultimate intention of this work is to apply the inference approach to the data on TB in badgers from Woodchester Park (Delahay et al., 2000). However, although the results in Section 4.3 look promising they are only for one group with high disease. Further testing is required to verify if it works on multiple groups with variable levels of disease. Therefore we do not include any results for the Woodchester Park data in this thesis. Even if we can demonstrate that the method works on simulated data from multiple groups, the Woodchester Park data will still present further challenges because the disease prevalence is generally much lower than in the simulated data we used in section 4.3. Low disease prevalence implies low numbers of disease events and as shown in Section 4.3 and in Chapter 3.3.6 we can not expect to obtain good estimates for a parameter if the number of events related to it is too low. In the remainder of this section we will discuss some of the assumptions we have made in our approach in relation to the processes that are thought to occur in badger populations and the observation and testing practices in Woodchester Park.

In comparison to published models for TB in badgers (Anderson and Trewhella (1985); White and Harris (1995a); Smith et al. (1997), see also Appendix A) the

process model presented here is very simple and this affects our estimates in a number of ways. As was discussed in Chapter 3 the absence of a demographic structure (age and gender) and seasonality means that the estimates from our model only give us annual per capita rates and can not directly be compared to those used in more complicated models that do include these features. The disease component of our model also makes some simplifications compared to the other models (Appendix A), many of which include disease induced mortality and pseudo vertical transmission (transmission from mother to offspring). In the disease component of our model we include background transmission in order to provide a mechanism for the invasion of disease into a previously disease free group. In most published models this type of disease invasion can occur through the interaction between groups in a spatial framework, but because our model is non-spatial this is not possible. The background transmission component in our model can also be seen as a mechanism for infection from the environment (e.g. contaminated pasture (Courtenay et al., 2006)) or other species (e.g. cattle and or deer (White et al., 2008)).

Our model is an SEI model and as such also differs from many other models for TB in badgers. Apart from some of the early modelling papers (Bentil and Murray, 1993; Ruxton, 1996) it is generally agreed that badgers do not recover from TB. Later models (e.g. White and Harris (1995a); Smith et al. (1997); Shirley et al. (2003)) do include a recovered class R, but only because they consider vaccination. More importantly Smith et al. (1997) and Shirley et al. (2003) also introduced a second category of more highly infectious animals. These models are referred to as $SEII$ or $SEIIR$ models. The main reason for not including the additional category is that while implementing and testing the feasibility of the method the model had to remain simple, however, in principle the inference methods described in this thesis are applicable to more complex models such as these. Clifton-Hadley et al. (1993) suggested that badgers only shed TB intermittently and thus the straight forward progression from exposed to infectious (to highly infectious) is a simplification.

It is clear from the descriptions in Section 4.2.2 of the simulated data and the observations from Woodchester Park that there are some important differences between the two that also need to be considered when applying the inference approach. One difference is that in the simulation model we observe the system at one fixed point in the year whereas in Woodchester Park there are four field campaigns each year. One implication of this is that in our model we can not count individuals at the end of a period between observations if it died before the observation point. In contrast, in Woodchester Park, the four field campaigns per year mean that if for example an animal is caught in the spring or summer but dies before the end of the year it will be included in the count for the population size for that year. Similarly in our model we can only apply the disease tests at fixed time points and we can not take into account individuals that tested positive but died before the time of the next observation point.

The disease classification used by FERA in Woodchester Park assumes that once an individual tests positive in the ELISA test it is classified as Exposed E and remains counted as such until it is classified as Infectious I by testing positive in a culture test (Delahay et al., 2001). Obviously these classifications

are individual based and we can not deal with this in our model as currently implemented in the inference approach where effectively we only test at the end of each year.

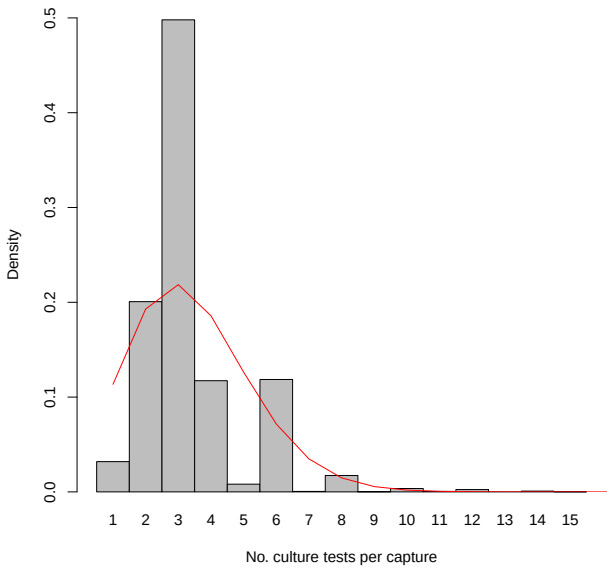


Figure 4.11 Analysis of the number of culture tests done on each captured individual recorded in the Woodchester Park database. This data can be fitted with a Poisson distribution with mean = 3.4 (red line).

Another consequence of using annual observations from Woodchester Park is that the detection probability in the observation model is an effective annual probability which is the result of four field campaigns. So if the probability of being observed during a particular campaign is p_c then the annual detection probability $p_d = 1 - (1 - p_c)^4$ and thus an annual detection probability of 80% corresponds to a capture probability in a single campaign of only about 33%. We can apply a similar reasoning to the ELISA test where there is usually one test per capture. It is generally thought that the sensitivity of a single ELISA test is around 40% (Clifton-Hadley et al., 1995; Delahay et al., 2003; Greenwald et al., 2003) but because any individual can be caught and tested up to four times every year effectively the "annual sensitivity" could be around 87%. The sensitivity for a single culture test appears less well defined. Chambers et al. (2002) published a sensitivity for the culture test of 27.5 % based on the number of badgers testing positive in a culture test after having previously been classified as infectious. Although not clearly stated it appears that this sensitivity does not apply to a single culture test but to all the tests performed on a single capture which as illustrated in Figure 4.11 can vary widely. Given that badgers classified as infectious may only shed TB intermittently as suggested by Clifton-Hadley et al. (1993), the sensitivity reported by Chambers et al. (2002) could indicate that badgers, even when they are classified as infectious I, are only able to pass on the disease in about a quarter of the time. Ideally all of these considerations

should be taken into account in the application of the inference approach to the data from Woodchester Park but this is beyond the scope of this work.

Chapter 5

Discussion and conclusions

This thesis demonstrates the value of stochastic process-based modelling of complex biological systems. In Chapter 2 we showed how such models can be used to understand the qualitative impact of fundamental processes (competition between individuals for increased fecundity) on disease dynamics (in particular disease persistence). Chapters 3 and 4 showed how dynamic process-based models can be integrated with data in order to infer parameter values and the uncertainty associated with such estimates enabling such models to be used for quantitative prediction. If the model structure adequately captures the nature of the relevant processes these predictions should be more robust than standard statistical models, ultimately enabling extrapolation beyond the domain of the data used to fit the model. In Chapter 3 we use a simple model for population dynamics and a generic Bayesian inference approach that takes into account observation error in order to quantify three key population dynamics parameters and their uncertainty. For the first time we are able to present such estimates based on the long term observations for badgers in Woodchester Park. In Chapter 4 we show for the first time using simulated data that it is feasible to simultaneously estimate population and disease parameters from observations that are affected by multiple sources of error.

5.1 Modelling differential fecundity

In Chapter 2 we present a generic model to illustrate how differential fecundity results in demographic structures that can have a profound effect on disease persistence. This work adds to an extensive literature concerned with the effect of various types of population heterogeneity on disease dynamics (e.g. Anderson and May (1984); Anderson and Trewhella (1985); Renshaw (1991); Tilman and Kareiva (1997); Gudelj and White (2004)). Recent modelling work focuses on defining and modeling complex contact networks (e.g. Keeling (1999); Read and Keeling (2003); Vazquez (2007); Kiss et al. (2008)) but in reality, full knowledge of contact networks is rarely available and as a result they are difficult to parameterise. Another drawback of such models is that they do not capture the dynamic nature of contact networks as present in many wildlife populations. Consequently, we take a mechanistic approach (Chapter 2), focusing on the dynamic

nature of the contact networks by modeling the processes of movement and dispersal within a structured population driven by fecundity differences. Building on the non-spatial model of Davidson et al. (2008), we embed the social structure in a spatial context. This takes into account the observation that individuals can improve their chances of reproductive success by moving to a rank with higher fecundity within their current group or in a group in another spatial location (Creel and Rabenold, 1994; Woodroffe et al., 1995). Although in our model (Chapter 2), the contact network is not defined explicitly, we argue that contact networks are at best an approximation to the contact processes occurring within a population (i.e. the processes in which transmission of disease can occur). Moreover, we emphasise that our model does capture some of the key characteristics of networks observed in animal populations (e.g. Cross et al. (2004); Lusseau and Newman (2004); Ramos-Fernández et al. (2009)). In particular, by restricting transmission to ranks only we ensure that the possible number of contacts between individuals is limited and highly clustered, while allowing individuals to move between ranks and groups may replace the need to explicitly include "small world" properties in the disease transmission network.

We demonstrate the flexibility of our modeling framework and its ability to represent a range of plausible population structures (Figure 2.1) as reflected in the distribution of the number of individuals in each rank as well as the distribution of the ages of individuals across ranks (subgroups with specific fecundity)(Figure 2.3). The concentration of older (mature) individuals in some ranks causes higher prevalence in those ranks and greater persistence in the entire group. The simulations presented in Chapter 2 demonstrate that this effect is strongest when disease transmission is restricted to ranks of equal fecundity, but is also present when the contact network is widened by including contacts with other ranks, pseudo-vertical transmission and to some extend transmission on movement attempts. The addition of spatial heterogeneity further amplifies the effect of differential fecundity and our simulations suggest that the number of ranks determines the relative magnitude of the fecundity effects compared to that of the purely spatial effects and that the fecundity effect becomes increasingly important as the number of ranks increases (Figure 2.6).

Chapter 2 is an example of how without modelling specific observational data, relatively complex models make a useful contribution in improving our understanding of ecological and epidemiological processes. Our results suggest the importance of understanding the structures that are common in wildlife populations and the mechanisms that drive them. This is particularly important when we use models to examine the impact of disease control strategies in structured wildlife populations.

5.2 Inference approach

In Chapters 3 and 4 we argue that, in populations that are subject to significant size fluctuations, population and disease dynamics parameters should be inferred simultaneously. In Chapter 4 we demonstrate using simulated data that this is feasible and our results also suggests that in the presence of uncertainty with

respect to the observation parameters (p_d, P_E and P_C) this should also be included in the inference (Figure 4.10).

We have shown that increasing the amount of data available for parameter estimation affects the estimates of the process and observation parameters in different ways. The observation parameters benefit most from increasing the observation frequency (Figure 3.8 and Section 3.3.6). The estimates of process parameters also benefit from more frequently observed data, however, *more* is gained by observing for longer or by observing the same process multiple times as both of these strategies increase the number of events happening within the observation period. It is clear that this has implications for experimental design. Optimal design depends on whether the focus is on understanding the processes affecting a system (e.g. birth and death rates) or on the estimation of the state space variables (i.e. population size or number in the different disease classes). We show that this conclusion can be directly obtained from mathematically analysing the posterior distributions (Section 3.3.6)

As was noted in Section 4.4 the level of disease in the Woodchester Park data is relatively low (compared to the simulated data) and consequently does not include many disease related events. We are optimistic that by combining the likelihood of multiple simultaneously observed groups we should be able to estimate the disease parameters with this data set, but this has not yet been done.

5.3 Parameter heterogeneity in the inference approach

In Chapter 3 we discuss the need to allow parameters in the process model to vary between simultaneously observed groups. In particular, we implement a hierarchical prior for the carrying capacity because in reality (e.g. Woodchester Park) the carrying capacity depends on the size and quality of the territory occupied by a group of wild animals (this is particularly clear in the case of badgers, e.g. Figure 3.1(a)). Before applying the inference approach to the Woodchester Park Data, we demonstrate that using a hierarchical prior for K results in better estimates when inferring parameters using simulated data for multiple groups with varying carrying capacity. The other population dynamics parameters (i.e. birth and death rate) could also vary but these are thought to be intrinsically linked to a species and less influenced by inter group variability linked to environmental factors. If such variability is present it will in any case be included in the uncertainty in the estimate. Our approach further assumes that there is no variability over time and therefore any seasonal variability is absorbed in the parameter estimates. Any long term fluctuations linked to climatic or environmental conditions will be reflected in the parameter uncertainty.

In the disease component of the model currently implemented we have assumed that there is no parameter heterogeneity between groups. This seems reasonable for the disease transmission rate β and the latency rate τ_{EI} which can be thought to be intrinsically linked to a specific disease (e.g. Tuberculosis).

However it was suggested by Clifton-Hadley et al. (1993) that the intermittent shedding of TB in badgers could be caused by external stress factors (e.g. food shortage), which could affect some groups more than others and thus result in a variation in the transmission rate. The background transmission rate is more closely linked to the environment and is therefore quite likely to vary between groups.

One of the main advances to the inference approach presented in this thesis is the estimation of the observation parameters. These are considered not to vary between groups or over time. Obviously this is a simplification as the effort invested in the observations may vary between groups (resulting in varying capture rates) and over time. Similarly we have assumed no variation in the disease test sensitivities. We assume that all individuals are tested in the same way which obviously is not true in the case of the culture test (Figure 4.11) applied to badgers in Woodchester Park. Furthermore we make the implicit assumption that the test sensitivity is constant over time, i.e. the same test is used in the same way throughout the observation period. This is a valid assumption for the observations from Woodchester Park, although the possible introduction of new tests (e.g. Dalley et al. (2008)) will have to be accounted for if such data were to be used.

There is no theoretical limitation to the number of parameters that could be allowed to vary in the inference approach. However, in practical terms the implementation of hierarchical priors leads to complications in coding of the algorithm, larger data output and longer run times. In the case of the carrying capacity, in each iteration instead of a taking single sample of K, we sample and output the mean and standard deviation of K separately as well as all the group carrying capacities K_i in each of the groups, which results in a significantly increased computational burden.

5.4 General applicability of the inference approach

The approach, as currently implemented, can in principle be applied to any long term observations for population and/or disease dynamics and possible candidates include the observations of badgers in Wytham Woods (Macdonald and Newman, 2002; Macdonald et al., 2009), voles in Kielder Forest (Telfer et al., 2007; Smith et al., 2009) and Soay sheep (e.g. Milner et al. (2001)). The application does not have to be restricted to wild animals and as an example could also be used on data from epidemiological observations on livestock (e.g. (Höhle et al., 2005)). As we have demonstrated in Chapters 3 and 4 there are limitations that determine how much can be learned from a particular data set. The main limitation for the process parameters is the number of events included in the observation period (Chapter 3) and therefore in general the method will perform better if a system is observed for longer or simultaneously on multiple groups (Gibson et al., 2006). It could be argued that it is also better to observe larger populations as they will have larger total event rates. In the case of our analysis of the Woodchester Park

data we considered treating the population as a whole but we did not pursue this because population and disease dynamics processes are likely to be mainly controlled by local factors with limited interaction between groups (e.g. migration and intergroup disease transmission).

Having derived parameter estimates including their associated uncertainty we should aim to use these in a predictive context but this is beyond the scope of this thesis. An example of where our results could be particularly useful is in the context of the Randomised Badger Culling Trial (ISG, 2007) which proved that culling badgers had the effect of increasing the level of TB in cattle as well as the badger population (Jenkins et al., 2007). This is referred to as the "perturbation effect" and was not predicted by any of the existing models (e.g. (White and Harris, 1995b; Swinton et al., 1997; Smith et al., 1997)). Recent work by Prentice (2012) showed that under specific conditions the perturbation effect can result from very simple models without explicitly changing the parameters at the point of culling. Obviously including the parameter estimates from this study would allow some measure of uncertainty to be given for the outcomes Prentice's work.

5.5 Future work

There are several ways in which the work in this thesis could be extended.

- The modelling and simulation work in Chapter 2 is based on general concepts and ideas from ecology and epidemiology and demonstrates that differential fecundity can have a significant effect on disease prevalence and persistance. This effect has sometimes been hinted at in natural systems, e.g. in rabbits (Judge et al., 2006) but the value of the work in this thesis would greatly increase if this effect could be demonstrated in observations. In that case it is essential that all components of the system are observed, including the structures that are included in the model, how these affect the transmission routes and the resulting levels of disease.

- Understanding of the model outcomes in Chapter 2 could be developed further by investigating the contact network in the model and quantifying the basic reproduction number (R_0), which is commonly used in epidemiology to characterise the strength of an epidemic.

- The models used in Chapters 3 and 4 are by necessity simple. It could be usefull to investigate if the approach can be applied to more complex models, e.g. by including age and gender or spatial interaction in the population dynamics model.

- In the inference approach in Chapters 3 and 4 we do not include the individual based information that is present in the Woodchester Park database. More complete use of this type of information could be considered as may affect the estimates of some of the model parameters.

- Having inferred population dynamics parameters and their uncertainty for the Woodchester Park data in Chapter 3, these should now be used in simulation. The results from such simulations could then reflect the uncertainty in the parameters that is demonstrated in this thesis.

- The inference method (Chapter 4) needs to be demonstrated to work with the low disease prevalence and very low numbers of positive in the disease test as observed in the Woodchester Park study.

- In the case of the inference on the population dynamics parameters in Chapter 3, it was shown that using multiple groups yielded estimates that were less dependent on the prior. Extending this to the inference method including disease (Chapter 4) could also be investigated.

5.6 Conclusions

The main conclusions in this thesis are:

- Using a model for demography and disease dynamics we show that differential fecundity and the resulting population structure, as represented in the distribution of individuals and their ages across smaller sub-populations, have a profound impact on disease persistence.

- Including spatial heterogeneity by embedding local populations affected by differential fecundity in a larger spatial framework that includes interaction between such local population further amplifies the effect of differential fecundity on disease persistence.

- We demonstrate that the effect of differential fecundity on disease persistence is robust under a range of disease transmission scenarios.

- We implemented the inference approach by Marion et al. (2012) for stochastic continuous time discrete state space models and for the first time included a methodology to allow for the estimation of uncertain observation parameters.

- We demonstrated the feasibility of simultaneously estimating all parameters (r_b, μ, K and p_d) and their uncertainty for a simple population model with births and deaths only using annually observed data with significant observational error.

- The inferred state space history forms an integral part of our inference approach and provides an alternative method for inferring "true" population histories and their associated uncertainty.

- Including data from multiple groups in the inference by combining their likelihoods leads to a significant improvement in the parameter estimates.

- We explain this dependence by analysing the definition of the posterior distribution for a given *process parameter* which shows that the coefficient of variation of the posterior of the parameter decreases as a function of the number of events of the relevant type (e.g. births or deaths) that occur within the observation period.

- The frequency with which we observe the state space variables has an indirect effect on the parameter estimates because the difference between states in subsequent observations carries implicit constraints on the number of events between subsequent observations.

- Using simulated data we show that estimation of the detection probability p_d for a population model is feasible but it depends strongly on the frequency of observation.

- We explain this improvement by analysing the posterior distribution for the *observation parameters* which shows that the coefficient of variation for observation parameters decreases as a function of the number observations.

- The coefficient of variation for the observation parameters also depends on the size of the state being observed, e.g. the coefficient of variation for the detection probability decreases as the size of the observed population increases.

- We implemented a hierarchical prior in the population model to allow for between group variation in the carrying capacity K and demonstrated using simulated data with that this improves the estimate of the other parameters (r_b and μ).

- Application to the long term observations Woodchester Park data allows us for the first time to present estimates for the birth and death rates for this undisturbed badger population including uncertainty. The median and 95% Credible Interval (in brackets) for the birth rate r_b is $0.58(0.45, 0.74)$ and the for the death rate μ it is $0.40(0.31, 0.51)$.

- We implemented the inference approach for a combined population and disease dynamics model and with an observation model that includes both a capture process as well as two disease tests (ELISA and culture tests).

- We demonstrate, using simulated data for a single group, that simultaneous estimation of all parameters (i.e. population, disease and observation parameters) is feasible if disease prevalence is high.

- We make several simplifications in the definition of our model with respect to real ecological processes, field observations and disease testing practices and discuss this in the specific context of the data on TB in badgers from Woodchester Park. In spite of these limitations, our approach represents considerable progress towards simultaneous estimation of population and disease dynamics parameters in the presence of significant observation errors for data from natural systems.

Bibliography

S. Altizer, C.L. Nunn, P.H. Thrall, J.L. Gittleman, J. Antonovics, A.A. Cunningham, A.P. Dobson, V. Ezenwa, K.E. Jones, A.B. Pedersen, M. Poss, and R.C. Pulliam. Social Organization and Parasite Risk in Mammals: Integrating Theory and Empirical Studies. *Annual Review of Ecology, Evolution, and Systematics*, 34(1):517–547, 2003.

R.M. Anderson and R.M. May. Population biology of infectious diseases: Part i. *Nature*, 280(5721):361–367, 1979.

R.M. Anderson and R.M. May. Spatial, temporal, and genetic heterogeneity in host populations and the design of immunization programmes. *Mathematical Medicine and Biology*, 1(3):233–266, 1984.

R.M. Anderson and W. Trewhella. Population Dynamics of the Badger (*Meles meles*) and the Epidemiology of Bovine Tuberculosis (*Mycobacterium bovis*). *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences (1934-1990)*, 310(1145):327–381, 1985.

F. Ball. Stochastic and deterministic models for sis epidemics among a population partitioned into households. *Mathematical Biosciences*, 156(1-2):41–67, 1999.

N.D. Barlow. The ecology of wildlife disease control: Simple models revisited. *Journal of Applied Ecology*, 33(2):303–314, 1996.

D.E. Bentil and J.D. Murray. Modelling bovine tuberculosis in badgers. *Journal of Animal Ecology*, 62(2):239–250, 1993.

B. Bolker and S.W. Pacala. Using moment equations to understand stochastically driven spatial pattern formation in ecological systems. *Theoretical Population Biology*, 52(3):179–197, 1997.

R.J. Boys, D.J. Wilkinson, and T.B.L. Kirkwood. Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18(2): 125–135, 2008.

S.T. Buckland, K.B. Newman, C. Fernández, L. Thomas, and J. Harwood. Embedding population dynamics models in inference. *Statistical Science*, 22(1): 44–58, 2007.

S. Catterall, A.R. Cook, G. Marion, A. Butler, and P.E. Hulme. Accounting for uncertainty in colonisation times: a novel approach to modelling the spatio-temporal dynamics of alien invasions using distribution data. *Ecography*, 35: 901–911, 2012.

M A Chambers, W A Pressling, C L Cheeseman, R S Clifton-Hadley, and R G Hewinson. Value of existing serological tests for identifying badgers that shed *Mycobacterium bovis*. *Veterinary Microbiology*, 86(3):183–189, 2002.

E.C.J. Claas, A.D.M.E. Osterhaus, R. van Beek, J.C. De Jong, G.F. Rimmelzwaan, D.A. Senne, S. Krauss, K.F. Shortridge, and R.G. Webster. Human influenza a H5N1 virus related to a highly pathogenic avian influenza virus. *The Lancet*, 351(9101):472–477, 1998.

J.S. Clark and O.N. Bjørnstad. Population time series: process variability, observation errors, missing values, lags, and hidden states. *Ecology*, 85(11):3140–3150, 2004.

R.S. Clifton-Hadley, J.W. Wilesmith, and F.A. Stuart. Mycobacterium bovis in the european badger *(Meles meles)*: epidemiological findings in tuberculous badgers from a naturally infected population. *Epidemiology and Infection*, 111 (1):9–19, 1993.

R.S. Clifton-Hadley, A.R. Sayers, and M.P. Stock. Evaluation of an elisa for *Mycobacterium bovis* infection in badgers (*Meles meles*). *Veterinary Record*, 137(22):555–558, 1995.

O. Courtenay, L.A. Reilly, F.P. Sweeney, V. Hibberd, S. Bryan, A. Ul-Hassan, C. Newman, D.W. Macdonald, R.J. Delahay, G.J. Wilson, and E.M.H. Wellington. Is *Mycobacterium bovis* in the environment important for the persistence of bovine tuberculosis? *Biology Letters*, 2(3):460–462, 2006.

D.P. Cowan and P.J. Garson. *Behavioural Ecology: Ecological Consequences*, chapter Variations in the social structure of rabbit populations: Causes and demographic consequences., pages 537–555. Blackwell Scientific Publications, Oxford, 1985.

M.K. Cowles and B.P. Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434): 883–904, 1996.

D.R. Cox and H.D. Miller. *The theory of stochastic processes*. Wiley, New York, 1965.

S.R. Creel and K.N. Rabenold. Inclusive fitness and reproductive strategies in dwarf mongooses. *Behavioral Ecology*, 5(3):339–348, 1994.

P.C. Cross, J.O. Lloyd-Smith, J.A. Bowers, C.T. Hay, M. Hofmeyr, and W.M. Getz. Integrating association data and disease dynamics in a social ungulate: bovine tuberculosis in African buffalo in the Kruger National Park. In *Annales*

Zoologici Fennici, volume 41, pages 879–892. Finnish Zoological and Botanical Publishing Board Helsinki, Finland, 2004.

D. Dalley, D. Davé, S. Lesellier, S. Palmer, T. Crawshaw, R.G. Hewinson, and M. Chambers. Development and evaluation of a gamma-interferon assay for tuberculosis in badgers (*Meles meles*). *Tuberculosis*, 88(3):235–243, 2008.

P. Daszak, A.A. Cunningham, and A.D. Hyatt. Emerging infectious diseases of wildlife–threats to biodiversity and human health. *Science*, 287(5452):443–449, 2000.

R.S. Davidson, G. Marion, and M.R. Hutchings. Effects of social hierarchy on disease persistence. *Journal of Theoretical Biology*, 253:424–433, 2008.

RS Davidson, G. Marion, PCL White, and MR Hutchings. Use of host population reduction to control wildlife infection: rabbits and paratuberculosis. *Epidemiology and Infection*, 137:131–138, 2009.

R.S. Davidson, I.J. McKendrick, J.C. Wood, G. Marion, A. Greig, K. Stevenson, M. Sharp, and M.R. Hutchings. Accounting for uncertainty in model-based prevalence estimation: paratuberculosis control in dairy herds. *BMC Veterinary Research*, 8(1):159, 2012.

M.S. de Villiers, P.R.K. Richardson, and A.S. van Jaarsveld. Patterns of coalition formation and spatial association in a social carnivore, the African wild dog (*Lycaon pictus*). *Journal of Zoology*, 260(04):377–389, 2003.

S.L. Deem, W.B. Karesh, and W. Weisman. Putting theory into practice: wildlife health in conservation. *Conservation Biology*, 15(5):1224–1233, 2001.

DEFRA. Government strategic framework for the sustainable control of bovine tuberculosis (btb) in great britain, a sub-strategy of the animal health and welfare strategy for great britain. Technical report, DEFRA, 2004.

R.J. Delahay, S. Langton, G.C. Smith, R.S. Clifton-Hadley, and C.L. Cheeseman. The spatio-temporal distribution of *Mycobacterium bovis* (bovine tuberculosis) infection in a high-density badger population. *Journal of Animal Ecology*, 69 (3):428–441, 2000.

R.J. Delahay, J.A. Brown, P.J. Mallinson, P.D. Spyvee, D. Handoll, L.M. Rogers, and C.L. Cheeseman. The use of marked bait in studies of the territorial organization of the european badger (*Meles meles*). *Mammal Review*, 30(2): 73–87, 2001.

R.J. Delahay, G.J. Wilson, G.C. Smith, and C.L. Cheeseman. Vaccinating badgers (*Meles meles*) against *Mycobacterium bovis*: the ecological considerations. *The Veterinary Journal*, 166(1):43–51, 2003.

D.A. Dewsbury. Dominance rank, copulatory behavior, and differential reproduction. *Quarterly Review of Biology*, 57(2):135–159, 1982.

C.A. Donnelly, R. Woodroffe, D.R. Cox, F.J. Bourne, C.L. Cheeseman, R.S Clifton-Hadley, G. Wei, G. Gettinby, P. Gilks, H. Jenkins, W.T. Johnston, A.M. Le Fevre, J.P. McInerney, and W.I. Morrison. Positive and negative effects of widespread badger culling on tuberculosis in cattle. *Nature*, 439: 843–846, 2006.

A.W.F. Edwards. *Likelihood*. John Hopkins University Press, Baltimore, 1992.

L. Ellis. Dominance and reproductive success among nonhuman animals: a cross-species comparison. *Ethology and Sociobiology*, 16(4):257–333, 1995.

N.M. Ferguson, R.M. May, and R.M. Anderson. *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions*, chapter Measles: Persistence and synchronicity in disease dynamics, pages 137–157. Princeton Univ. Press, Princeton, NJ, 1997.

J.A.N. Filipe and M.M. Maule. Analytical methods for predicting the behaviour of population models with general spatial interactions. *Mathematical Biosciences*, 183(1):15–35, 2003.

A.E. Gelfand and A.F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410): 398–409, 1990.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

G.J. Gibson and E. Renshaw. Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology*, 15 (1):19–41, 1998.

G.J. Gibson and E. Renshaw. Inference for immigration-death processes with single and paired immigrants. *Inverse Problems*, 17:455–466, 2001.

G.J. Gibson, W Otten, J.A.N. Filipe, A. Cook, G. Marion, and C.A. Gilligan. Bayesian estimation for percolation models of disease spread in plant populations. *Statistics and Computing*, 16(4):391–402, 2006.

D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711, 1995.

R. Greenwald, J. Esfandiari, S. Lesellier, R. Houghton, J. Pollock, C. Aagaard, P. Andersen, R.G. Hewinson, M. Chambers, and K. Lyashchenko. Improved serodetection of *Mycobacterium bovis* infection in badgers (*Meles meles*) using multiantigen test formats. *Diagnostic Microbiology and Infectious Disease*, 46 (3):197–203, 2003.

I. Gudelj and K.A.J. White. Spatial heterogeneity, social structure and disease dynamics of animal populations. *Theoretical Population Biology*, 66(2):139–149, 2004.

W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

G. Hide, E.K. Morley, J.M. Hughes, O. Gerwash, M.S. Elmahaishi, K.H. Elmahaishi, D. Thomasson, E.A. Wright, R.H. Williams, R.G. Murphy, and J.E. Smith. Evidence for high levels of vertical transmission in Toxoplasma gondii. *Parasitology*, 136(14):1877–1885, 2009.

M. Höhle, E. Jørgensen, and P.D. O'Neill. Inference in disease transmission experiments by using stochastic epidemic models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(2):349–366, 2005.

C.B. Huffaker. Experimental studies on predation: dispersion factors and predator-prey oscillations. *Hilgardia*, 27:343–383, 1958.

F.A. Huntingford, A.K. Turner, and L.M. Downie. *Animal conflict.* Chapman & Hall/CRC, London, 1987.

T. Irony and N. Singpurwalla. Noninformative priors do not exist: A discussion with jose m. bernardo. *Journal of Statistical Inference and Planning*, 65(1): 159–189, 1997.

ISG. Bovine TB: the scientific evidence. a science base for a sustainable strategy to control TB in cattle. final report of the independent scientific group on cattle TB. Technical report, 2007.

H.E. Jenkins, R. Woodroff, C.A. Donnelly, D.R. Cox, W.T. Jonhston, F.J. Bourne, C.L. Cheeseman, R.S. Clifton-Hadley, G. Gettinby, P. Gilks, R.G. Hewison, J.P. McInerey, and W.I. Morrison. Effects of culling on spatial associations of *Mycobacterium bovis* infections in badgers and cattle. *Journal of Applied Ecology*, 44(5):897–908, 2007.

K.E. Jones, N.G. Patel, M.A. Levy, A. Storeygard, D. Balk, J.L. Gittleman, and P. Daszak. Global trends in emerging infectious diseases. *Nature*, 451(7181): 990–993, 2008.

J. Judge, I. Kyriazakis, A. Greig, R.S. Davidson, and M.R. Hutchings. Routes of intraspecies transmission of *Mycobacterium avium* subsp. paratuberculosis in rabbits (*Oryctolagus cuniculus*): a field study. *Applied and Environmental Microbiology*, 72(1):398–403, 2006.

J. Judge, R.S. Davidson, G. Marion, P.C.L. White, and M.R. Hutchings. Persistence of *Mycobacterium avium* subspecies paratuberculosis in rabbits: the interplay between horizontal and vertical transmission. *Journal of Applied Ecology*, 44(2):302–311, 2007.

R.E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

M. Keeling. The implications of network structure for epidemic dynamics. *Theoretical Population Biology*, 67:1–8, 2005.

M.J. Keeling. The effects of local spatial structure on epidemiological invasions. *Proceedings of the Royal Society B: Biological Sciences*, 266(1421):859–867, 1999.

M.J. Keeling. Metapopulation moments: coupling, stochasticity and persistence. *Journal of Animal Ecology*, 69:725–736, 2000.

M.J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton, N.J., 1st edition, 2007.

W.O. Kermack and A.G. McKendrick. Contributions to the mathematical theory of epidemics. *Proceedings of the Royal society of London. Series A*, 115(834): 700–721, 1927.

R. King, B.J.T. Morgan, O. Gimenez, and S.P. Brooks. *Bayesian analysis for population ecology*. Chapman & Hall, London, 2009.

I.Z. Kiss, D.M. Green, and R.R. Kao. The effect of network mixing patterns on epidemic dynamics and the efficacy of disease contact tracing. *Journal of the Royal Society Interface*, 5(24):791–799, 2008.

H. Kruuk. *The Social Badger: Ecology and Behaviour of a Group-living Carnivore (Meles Meles)*. Oxford University Press, 1989.

R.J. Kubiak and A.R. McLean. Why was the 2009 influenza pandemic in England so small? *PLoS One*, 7(2):e30223, 2012.

S. Kulkarni and P. Heeb. Social and sexual behaviours aid transmission of bacteria in birds. *Behavioural processes*, 74(1):88–92, 2007.

A.L. Lloyd and R.M. May. Spatial heterogeneity in epidemic models. *Journal of Theoretical Biology*, 179(1):1–11, 1996.

D. Lusseau and M.E.J. Newman. Identifying the role that animals play in their social networks. *Proceedings: Biological Sciences*, 271:477–481, 2004.

D.W. Macdonald and C. Newman. Population dynamics of badgers (*Meles meles*) in Oxfordshire, UK: numbers, density and cohort life histories, and a possible role of climate change in population growth. *Journal of Zoology*, 256(1):121–138, 2002.

D.W. Macdonald, C. Newman, P.M. Nouvellet, and C.D. Buesching. An analysis of eurasian badger (*Meles meles*) population dynamics: implications for regulatory mechanisms. *Journal of Mammalogy*, 90(6):1392–1403, 2009.

G. Marion, D.M. Walker, A. Cook, D.L. Swain, M.R. Hutchings, E. Charmley, J. Steel, and S. Coffey. Towards an integrated approach to stochastic process-based modelling: with applications to animal behaviour and spatio-temporal spread. *Redesigning animal agriculture: the challenge of the 21st Century*, pages 144–170, 2007.

G. Marion, G.J. McInerny, J. Pagel, S. Catterall, A.R. Cook, F. Hartig, and R.B. O'Hara. Parameter and uncertainty estimation for process-oriented population and distribution models: data, statistics and the niche. *Journal of Biogeography*, 2012.

G.J. McInerny and D.W. Purves. Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution*, 2(3):248–257, 2011.

J.W. McNutt. Sex-biased dispersal in african wild dogs, lycaon pictus. *Animal Behaviour*, 52(6):1067–1077, 1996.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, et al. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087, 1953.

J.M. Milner, D.A. Elston, and S.D. Albon. Estimating the contributions of population density and climatic fluctuations to interannual variation in survival of Soay sheep. *Journal of Animal Ecology*, 68(6):1235–1247, 2001.

K.B. Newman, C. Fernández, L. Thomas, and S.T. Buckland. Monte carlo inference for state–space models of wild animal populations. *Biometrics*, 65(2): 572–583, 2009.

P.D. O'Neill and G.O. Roberts. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 162:121–129, 1999.

J. Pagel and F.M. Schurr. Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography*, 21(2):293–304, 2012.

J. Prentice. *The perturbation effect in wildlife systems: An emergent property of simple models*. PhD thesis, University of York, 2012.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

G. Ramos-Fernández, D. Boyer, F. Aureli, and L.G. Vick. Association networks in spider monkeys (Ateles geoffroyi). *Behavioral Ecology and Sociobiology*, 63 (7):999–1013, 2009.

J.M. Read and M.J. Keeling. Disease evolution on networks: the role of contact structure. *Proceedings: Biological Sciences*, 270:699–708, 2003.

E. Renshaw. *Modelling biological populations in space and time*. Cambridge University Press, Cambridge, 1991.

E. Renshaw and G.J. Gibson. Can Markov chain Monte Carlo be usefully applied to stochastic processes with hidden birth times? *Inverse Problems*, 14:1581–1606, 1998.

H.G. Rödel and D. von Holst. Features of the early juvenile development predict competitive performance in male European rabbits. *Physiology & behavior*, 97 (3-4):495–502, 2009.

P. Rohani and G.D. Ruxton. Dispersal and stability in metapopulations. *Mathematical Medicine and Biology*, 16(3):297–306, 1999.

J.A. Royle and R.M. Dorazio. *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Academic Press, San Diego, CA, 2008.

G.D. Ruxton. The effects of stochasticity and seasonality on model dynamics: Bovine tuberculosis in badgers. *Journal of Animal Ecology*, 65(4):495–500, 1996.

M.D.F. Shirley, S.P. Rushton, G.C. Smith, A.B. South, and P.W.W. Lurz. Investigating the spatial dynamics of bovine tuberculosis in badger populations: evaluating an individual-based simulation model. *Ecological Modelling*, 167 (1-2):139–157, 2003.

G.C. Smith, M.S. Richards, R.S. Clifton-Hadley, and C.L. Cheeseman. Modelling bovine tuberculosis in badgers in England: preliminary results. *Mammalia(Paris)*, 59(4):639–650, 1995.

G.C. Smith, C.L. Cheeseman, and R.S. Clifton-Hadley. Modelling the control of bovine tuberculosis in badgers in England: Culling and the release of lactating females. *Journal of Applied Ecology*, 34(6):1375–1386, 1997.

G.C. Smith, C.L. Cheeseman, R.S. Clifton-Hadley, and D. Wilkinson. A model of bovine tuberculosis in the badger (*Meles meles*): an evaluation of control strategies. *Journal of Ap*, 38:509–519, 2001.

K.F. Smith, A.P. Dobson, F.E. McKenzie, L.A. Real, D.L. Smith, and M.L. Wilson. Ecological theory to enhance infectious disease control and public health policy. *Frontiers in Ecology and the Environment*, 3(1):29–37, 2005.

M.J. Smith, S. Telfer, E.R. Kallio, S. Burthe, A.R. Cook, X. Lambin, and M. Begon. Host–pathogen time series data in wildlife support a transmission function between density and frequency dependence. *Proceedings of the National Academy of Sciences*, 106(19):7905–7909, 2009.

J. Swinton, F. Tuyttens, D. Macdonald, D.J. Nokes, C.L. Cheeseman, and R. Clifton-Hadley. A comparison of fertility control and lethal control of

bovine tuberculosis in badgers: the impact of perturbation induced transmission. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 352:619–631, 1997.

S. Telfer, H.E. Clough, R.J. Birtles, M. Bennett, D. Carslake, S. Helyar, and M. Begon. Ecological differences and coexistence in a guild of microparasites: Bartonella in wild rodents. *Ecology*, 88(7):1841–1849, 2007.

D. Tilman and P.M. Kareiva. *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions*. Princeton University Press, Princeton, 1997.

A. Vazquez. Epidemic outbreaks on structured populations. *Journal of Theoretical Biology*, 245(1):125–129, 2007.

J. Vicente, R.J Delahay, N.J Walker, and C.L Cheeseman. Social organization and movement influence the incidence of bovine tuberculosis in an undisturbed high-density badger *Meles meles* population. *Journal of Animal Ecology*, 69: 815–828, 2007.

D. von Holst, H. Hutzelmeyer, P. Kaetzke, M. Khaschei, H.G. Rödel, and H. Schrutka. Social rank, fecundity and lifetime reproductive success in wild European rabbits (Oryctolagus cuniculus). *Behavioral Ecology and Sociobiology*, 51(3):245–254, 2002.

D.M. Walker, F.J. Pérez-Barbería, and G. Marion. Stochastic modelling of ecological processes using hybrid gibbs samplers. *Ecological modelling*, 198(1): 40–52, 2006.

P.C.L White and S. Harris. Bovine tuberculosis in badger (*Meles meles*) populations in Southwest England: the use of a spatial stochastic simulation model to understand the dynamics of the disease. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 349:391–413, 1995a.

P.C.L. White and S. Harris. Bovine tuberculosis in badger (*Meles meles*) populations in Southwest England: an assessment of past, present and possible future control strategies using simulation modelling. *Phil Trans R Soc Lond B*, 349: 415–432., 1995b.

P.C.L. White, GA Newton-Cross, M. Gray, R. Ashford, C. White, and G. Saunders. Spatial interactions and habitat use of rabbits on pasture and implications for the spread of rabbit haemorrhagic disease in New South Wales. *Wildlife Research*, 30(1):49–58, 2003.

P.C.L. White, M. Böhm, G. Marion, and M.R. Hutchings. Control of bovine tuberculosis in british livestock: there is no silver bullet. *Trends in Microbiology*, 16(9):420–427, 2008.

D.J. Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2):122–133, 2009.

HB Wilson and MP Hassell. Host–parasitoid spatial models: the interplay of demographic stochasticity and dynamics. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 264(1385):1189–1195, 1997.

R. Woodroffe, D.W. Macdonald, and J. Silva. Dispersal and philopatry in the European badger, *Meles meles*. *Journal of Zoology*, 237(2):227–239, 1995.

Appendix A

Summary of published models for population and disease dynamics for badgers and Bovine Tb

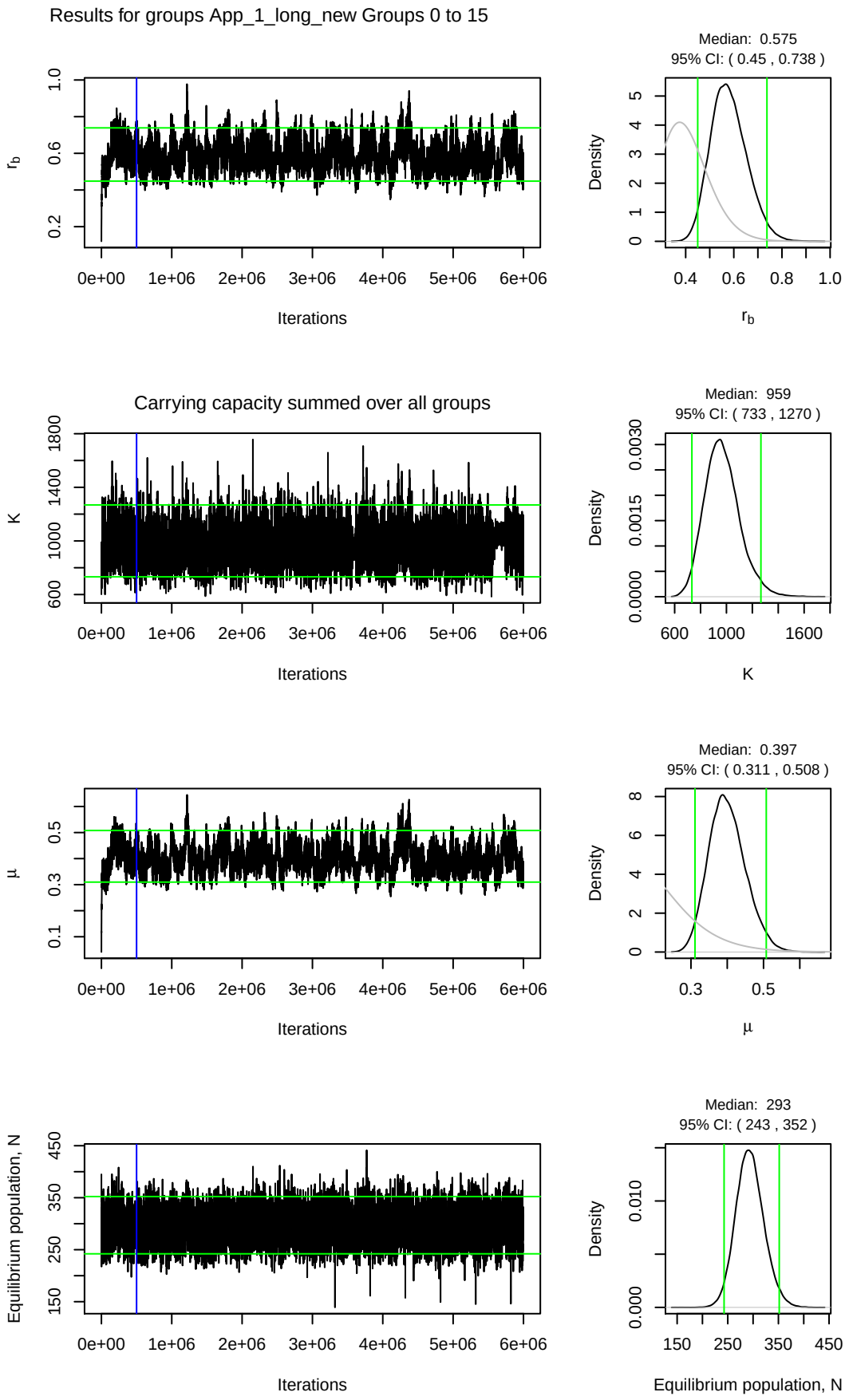
Comparison of published models for population and disease dynamics for badgers and Bovine Tb taken from Smith et al. (2001) and updated up to 2007.

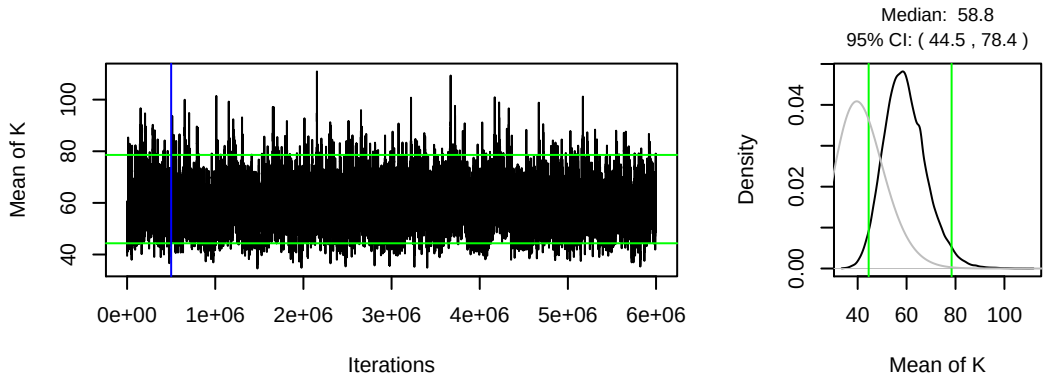
Key: Any gap indicates that the component was not included, 'y' indicates that the component was included, and 'c' that it was considered in some versions of the model. The model structure refers to the disease classification used. Where 'I' occurs twice differential infectivity is included. Note that some models include a Removed disease category, to represent those animals that have been vaccinated and as a result will never become infectious. Time is defined as 'D' discrete or 'C' continuous and the number of time steps within a year is included in seasonality for discrete time models. For age-structure, the number of categories is given. For isolation, 'i' indicates that the simulated population was completely isolated, 'f ' indicates that some form of disease-free immigration occurred, and 'b' indicates that a buffer population that contains TB was included. Disease transmission functions are described by 'd' for density-dependent, 'f ' for frequency-dependent, 'n' for any non-linear functions and 'l' for only local interactions. Spatial heterogeneity is either included in a random fashion (r) or as a specific geographic grid (s).

Model	Anderson and Trewhella (1995)	Bentil and Murray (1993)	Ruxton (1996)	Barlow (1996)	Swinton et al. (1997	Smith and Cheeseman (1999)	Whiet and Harris (1995a&b)	White et al. (1997)	Smith et al. (1995)	Smith et al. (1997)	Smith et al. ((2001a&b)	Shirley et al. (2003)	Wilkinson Smith et al. (2007) (2004)	DEFRA SE3108 (2004)
Structure	SEI	SEIR	SEIR	SEI	SI	SEI	SEIR*	SEIR*	SEII	SEIIR*	SEIIR*	SEIIR*	SEII	SEIIR*
Time	C	C	C	C	C	C	D	D	D	D	D	D	D	D
Spatial							Y	Y	Y	Y	Y	Y	Y	Y
Individual based												Y	Y	Y
Stochastic	C		Y				Y	Y	Y	Y	Y	Y	Y	Y
Seasonality	C		Y				4	4	1	1	2	2	2	2
Birth rate	D	L												
Age-structure	C	C2					3	3	6	6	3	3	3	3
Pseudo-vertical transmission	C								C	Y	Y	Y	Y?	Y
Habitat heterogeneity											r	s	R	S
Isolation	I	I	I	I	I	I	B	B	I	I	I& B	?	I	
Transmission	D	D	D	D,N	D,F	D	L	L	L	L	L	L?	L	L
Disease-induced mortality	1	0	0	1	0	0.22	0.63	0.48	I1=0.05 I2=0.6	I1=0.05 I2=0.6	I2=0.224	I2f=0.175 I2m=0.308	I2f=0.175 I2m=0.308	I2f=0.175 I2m=0.308
Dispersal							Y	Y	Y	Y	Y	Y	Y	Y
Cattle included											Y		Y	
Culling				Y	Y	Y	Y	Y		Y	Y		Y	
Vaccination				Y	Y	Y					Y		Y	
Fertility control				Y	Y	Y		Y						
Cost-benefit analysis													Y	

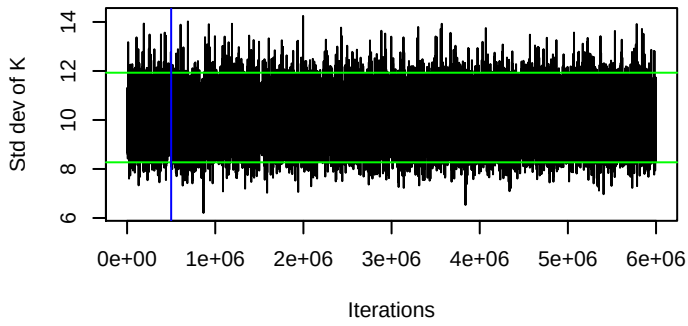
Appendix B

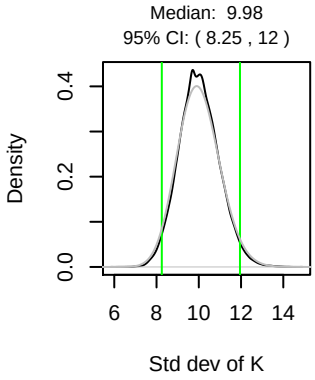
Full inference results for demographic parameters using Woodchester Park data

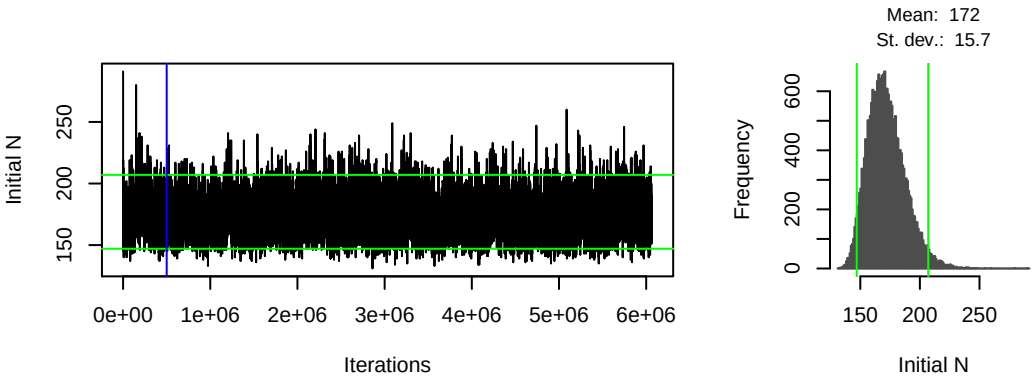

Results for groups App_1_long_new Groups 0 to 15
Median: 0.575
95% CI: (0.45 , 0.738)
r_b
Density
Iterations
Carrying capacity summed over all groups
Median: 959
95% CI: (733 , 1270)
K
Median: 0.397
95% CI: (0.311 , 0.508)
μ
Equilibrium population, N
Median: 293
95% CI: (243 , 352)

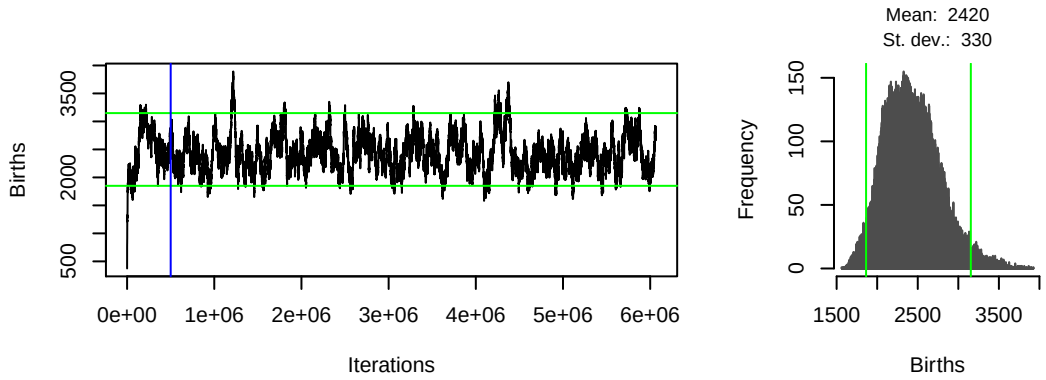

Mean of K
40
60
80
100
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
Iterations

Median: 58.8
95% CI: (44.5 , 78.4)
Density
0.00
0.02
0.04
40
60
80
100
Mean of K

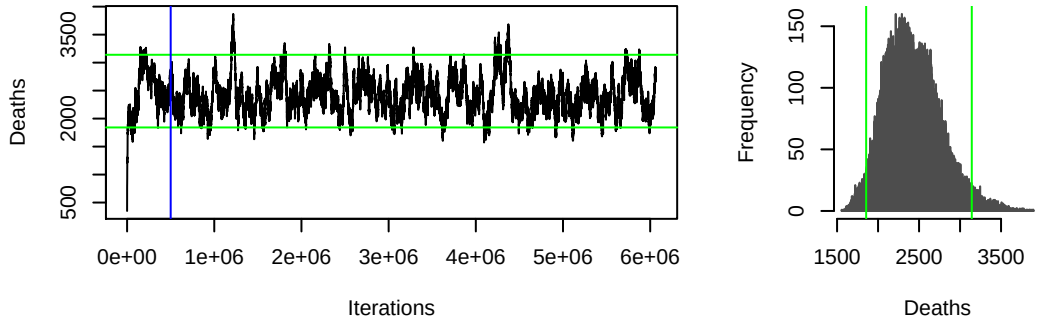

Std dev of K
6
8
10
12
14
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
Iterations


Median: 9.98
95% CI: (8.25 , 12)
Density
0.0
0.2
0.4
6
8
10
12
14
Std dev of K

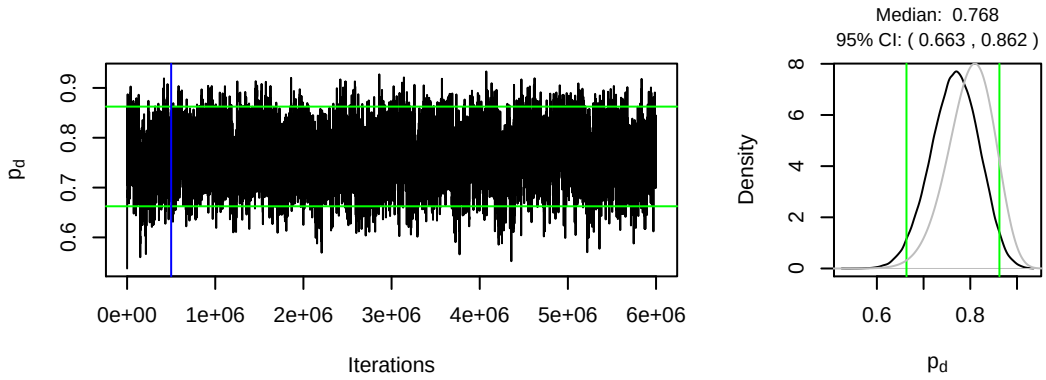

Initial N
150
200
250
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
Iterations
Mean: 172
St. dev.: 15.7
Frequency
0
200
400
600
150
200
250
Initial N

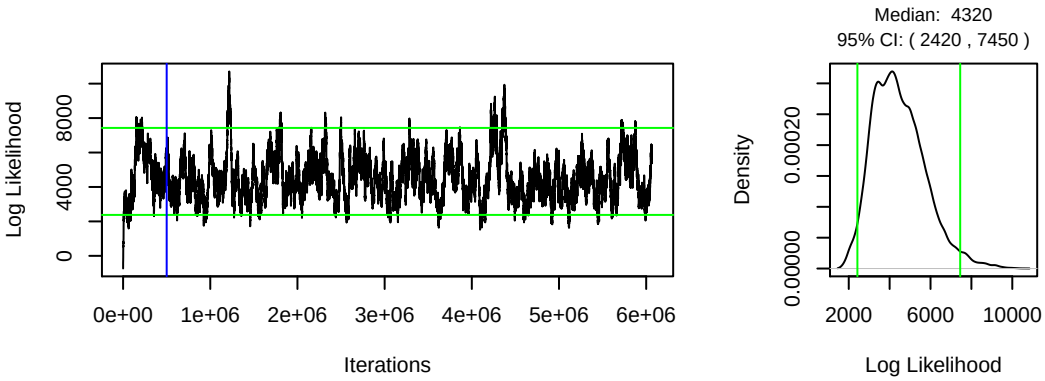

Births
500
2000
3500
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
Iterations

Mean: 2420
St. dev.: 330
Frequency
0
50
100
150
1500
2500
3500
Births

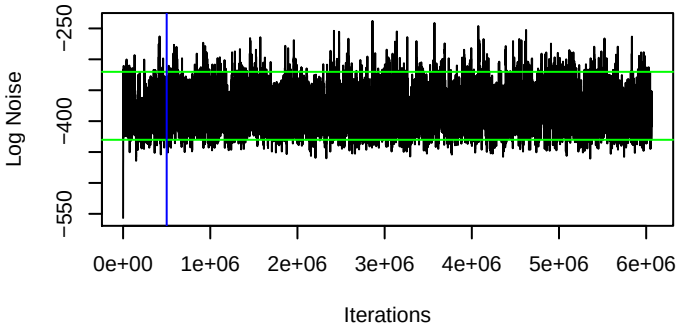

Deaths
500
2000
3500
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
Iterations

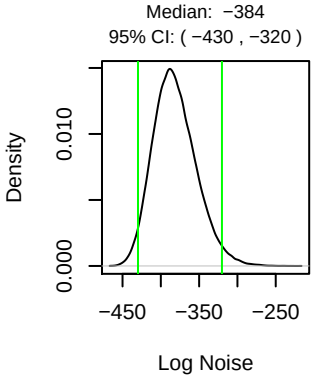
Mean: 2410
St. dev.: 329
Frequency
0
50
100
150
1500
2500
3500
Deaths


p_d
0.6
0.7
0.8
0.9
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
Iterations
Median: 0.768
95% CI: (0.663 , 0.862)
Density
0
2
4
6
8
0.6
0.8
p_d


Log Likelihood
0
4000
8000
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
Iterations

Median: 4320
95% CI: (2420 , 7450)
Density
0.00000
0.00020
2000
6000
10000
Log Likelihood

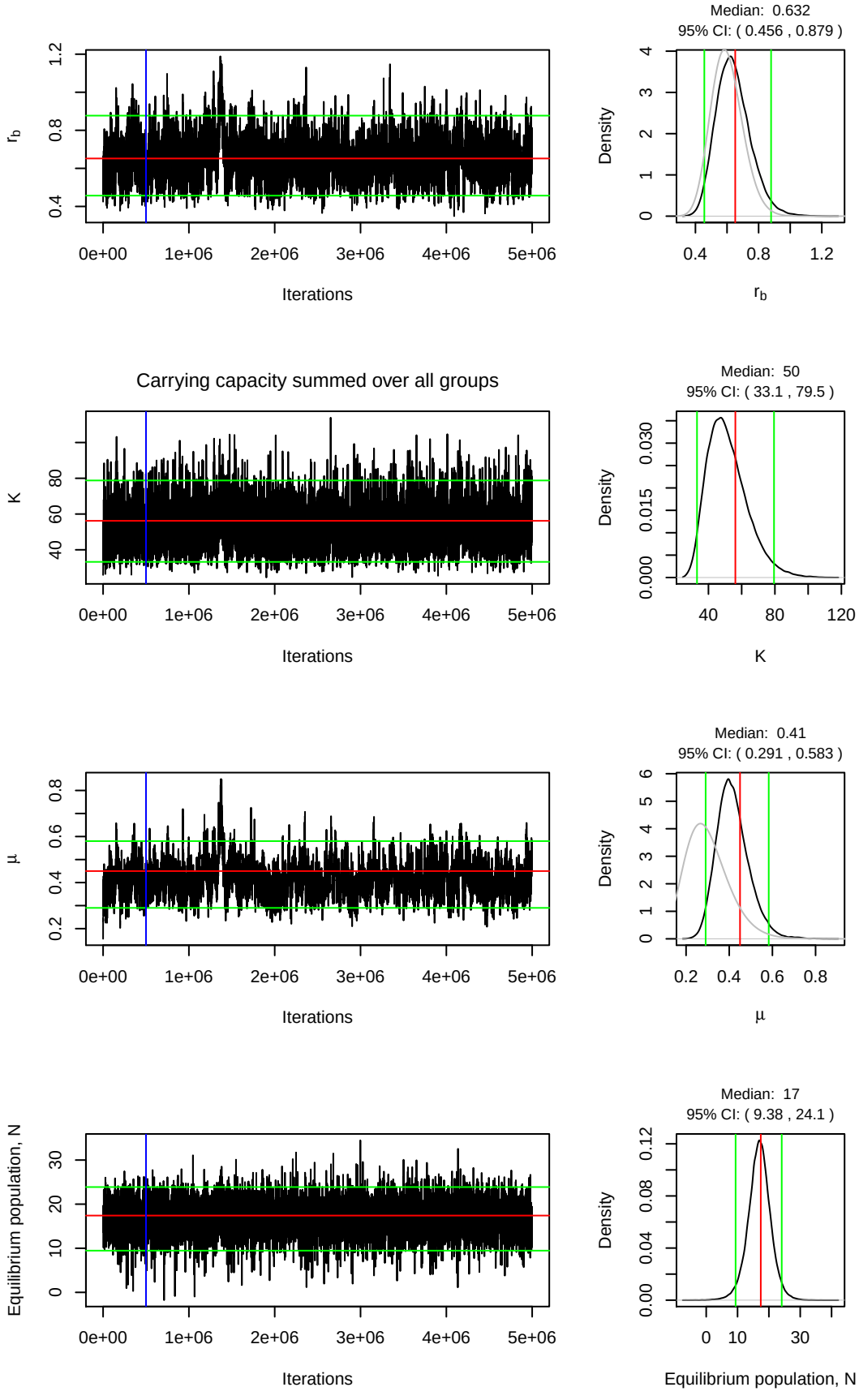

Log Noise
−250
−400
−550
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
Iterations

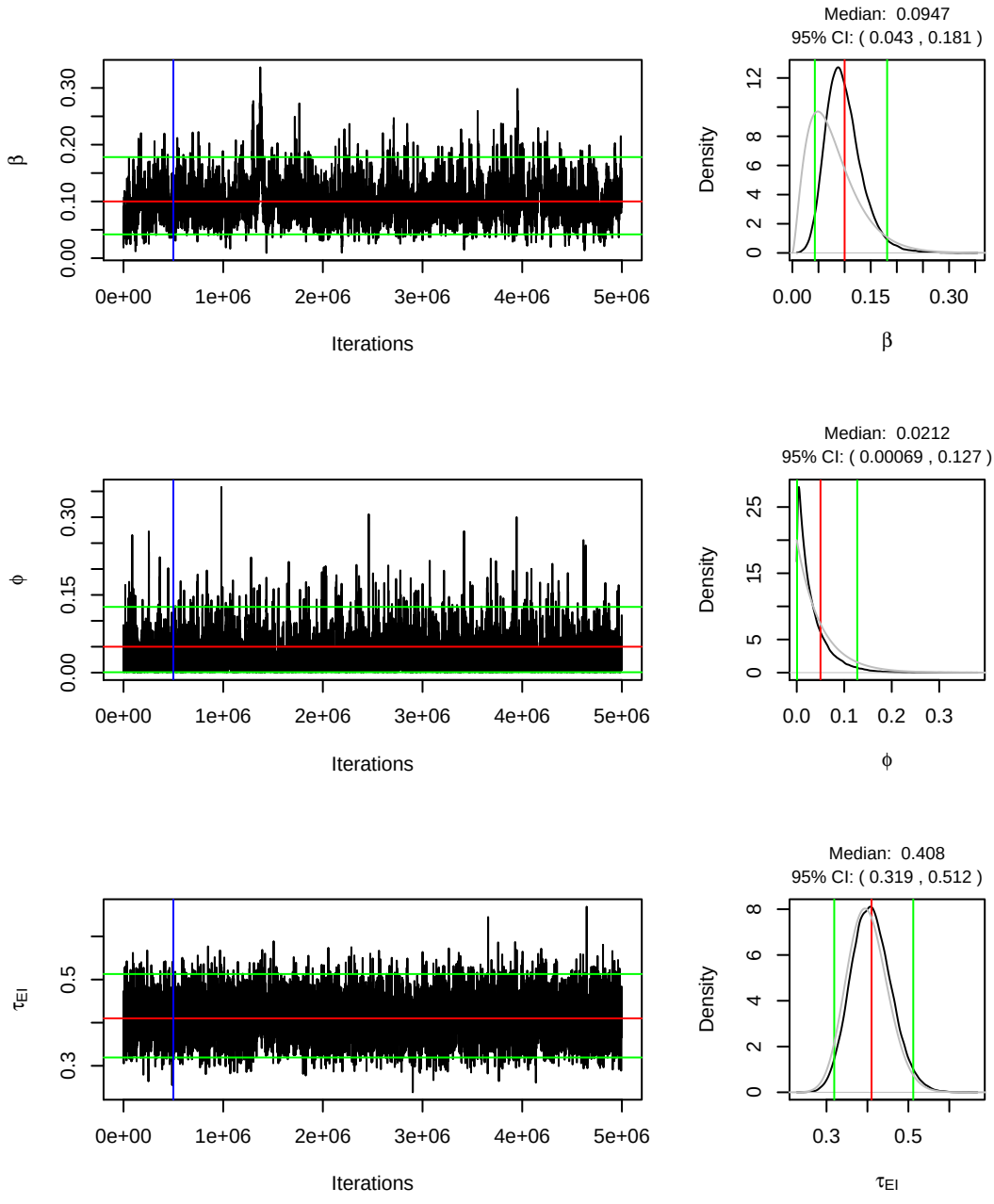

Median: −384
95% CI: (−430 , −320)
Density
0.000
0.010
−450
−350
−250
Log Noise

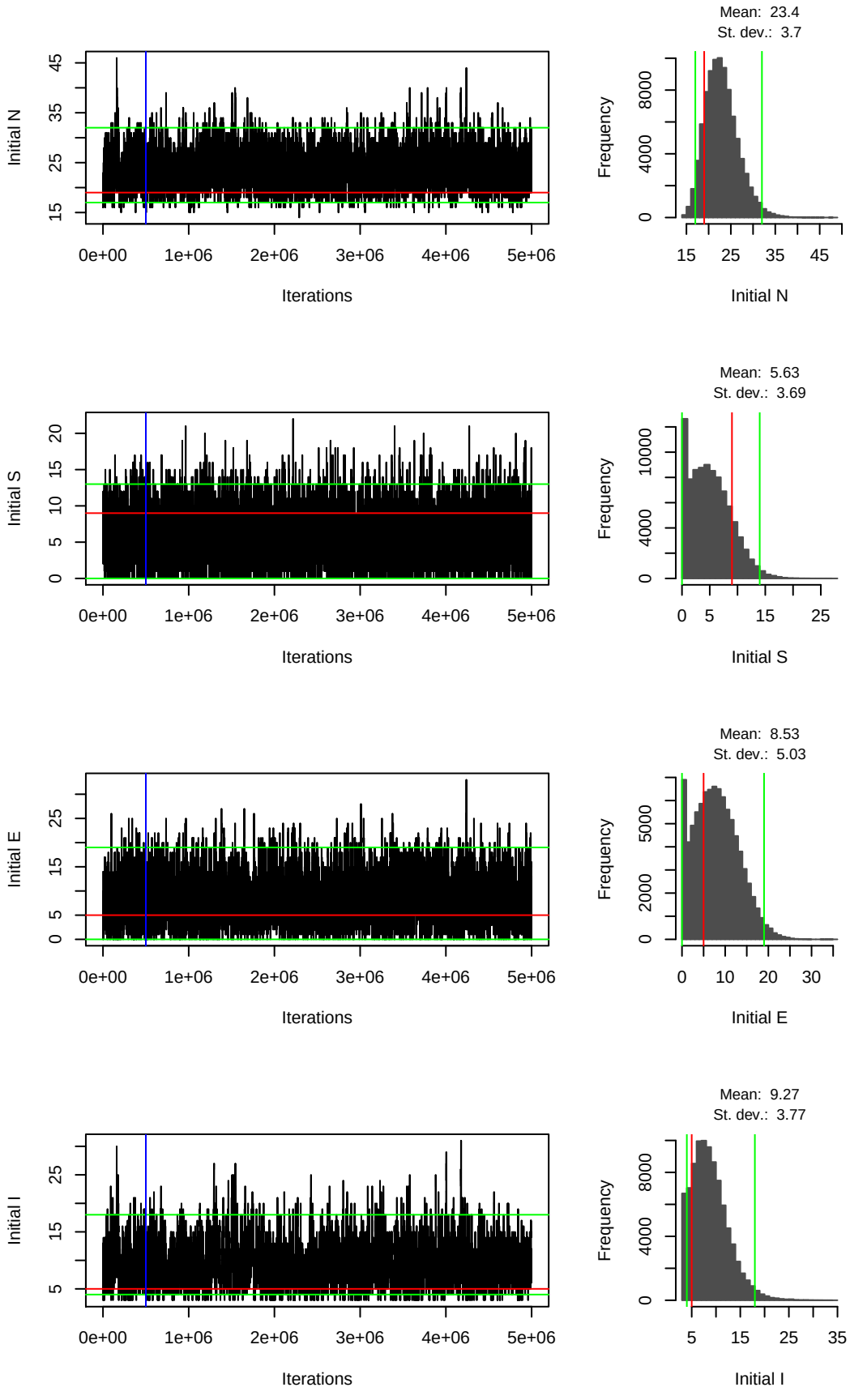
Appendix C

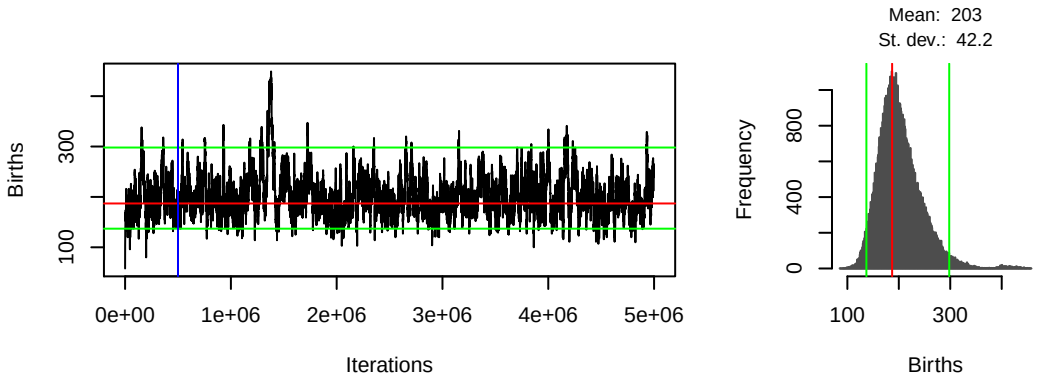
Full inference results for population and disease dynamics parameters using simulated data from a single group

Results for App_2_long Group 4

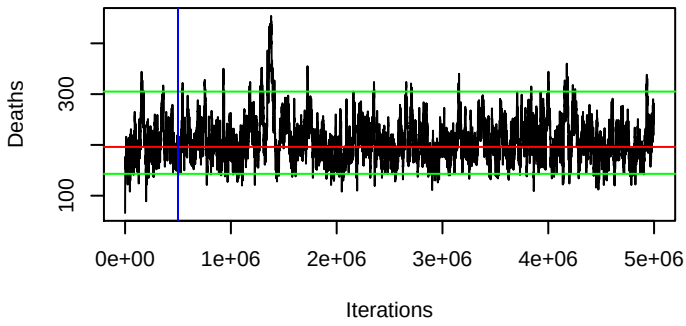


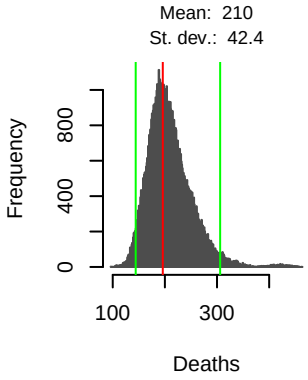

β
0.00
0.10
0.20
0.30
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Iterations
Median: 0.0947
95% CI: (0.043 , 0.181)
Density
0
2
4
6
8
12
0.00
0.15
0.30
β
ϕ
0.00
0.15
0.30
Iterations
Median: 0.0212
95% CI: (0.00069 , 0.127)
Density
0
5
15
25
0.0
0.1
0.2
0.3
ϕ
τEI
0.3
0.5
Iterations
Median: 0.408
95% CI: (0.319 , 0.512)
Density
0
2
4
6
8
0.3
0.5
τEI

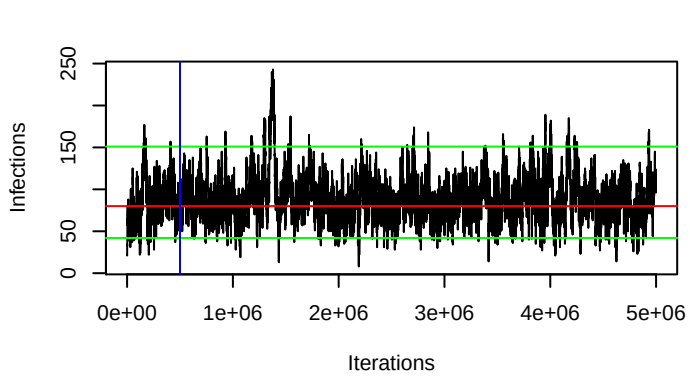

Initial N
Iterations
Mean: 23.4
St. dev.: 3.7
Frequency
Initial N
Initial S
Iterations
Mean: 5.63
St. dev.: 3.69
Frequency
Initial S
Initial E
Iterations
Mean: 8.53
St. dev.: 5.03
Frequency
Initial E
Initial I
Iterations
Mean: 9.27
St. dev.: 3.77
Frequency
Initial I

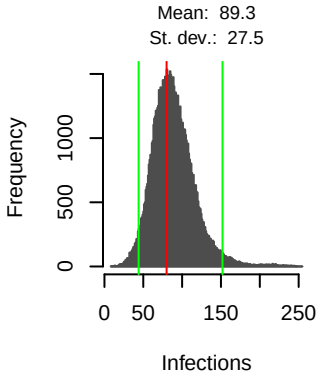

Births
100
300
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Iterations

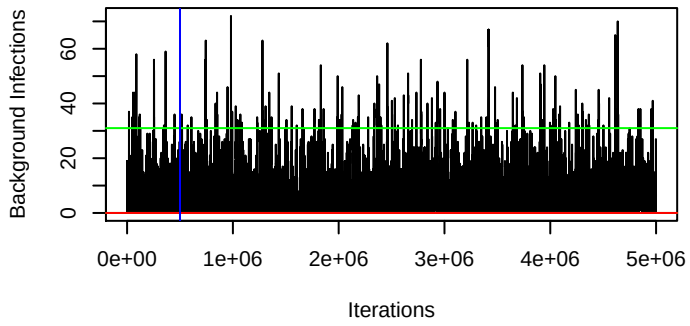
Mean: 203
St. dev.: 42.2
Frequency
0
400
800
100
300
Births

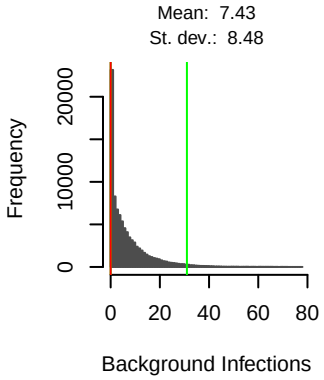

Deaths
100
300
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Iterations

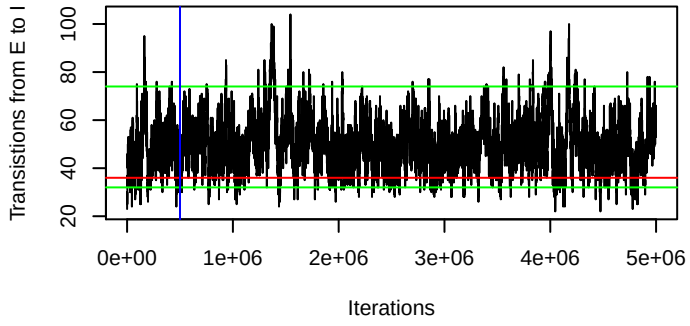

Mean: 210
St. dev.: 42.4
Frequency
0
400
800
100
300
Deaths

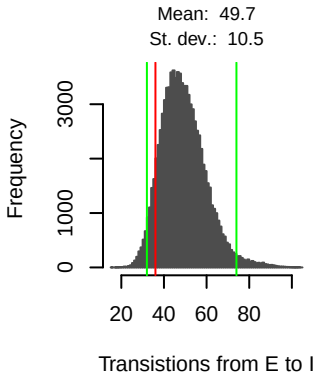

Infections
0
50
150
250
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Iterations

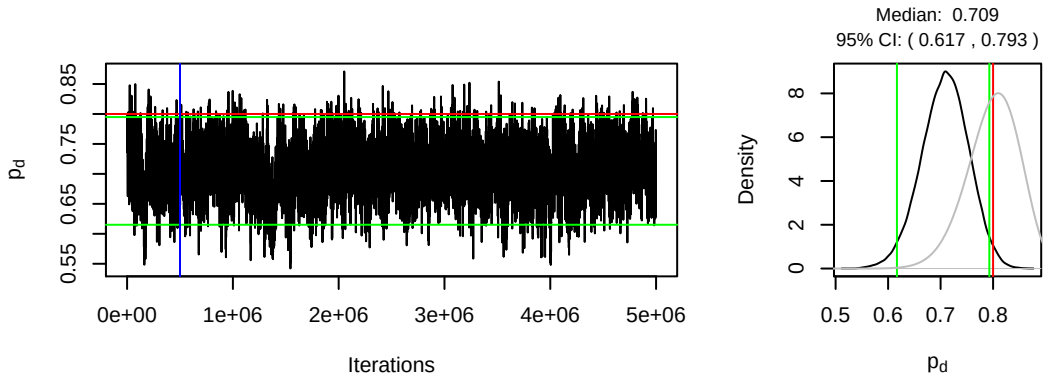

Mean: 89.3
St. dev.: 27.5
Frequency
0
500
1000
0
50
150
250
Infections

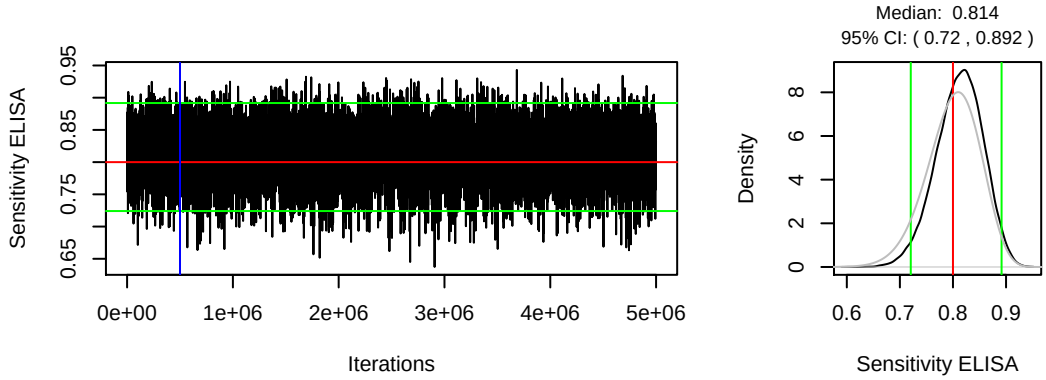

Background Infections
0
20
40
60
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Iterations

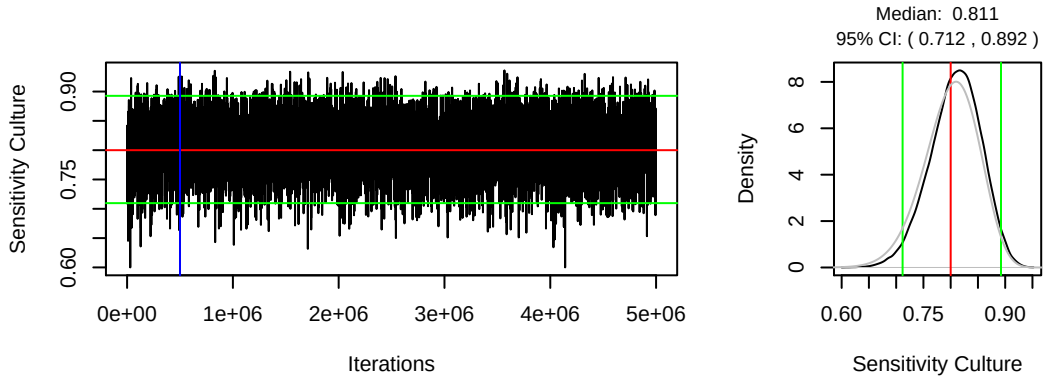

Mean: 7.43
St. dev.: 8.48
Frequency
0
10000
20000
0
20
40
60
80
Background Infections

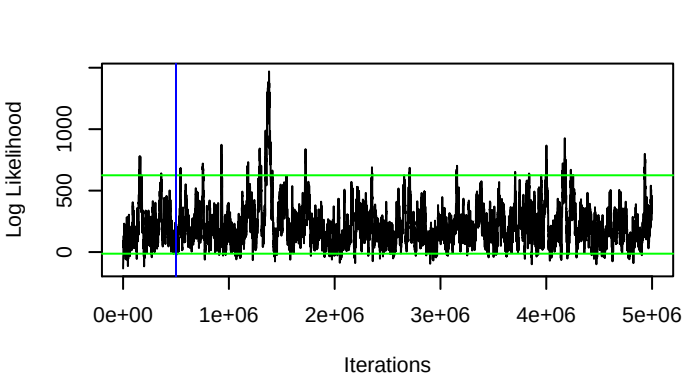

Transistions from E to I
20
40
60
80
100
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Iterations

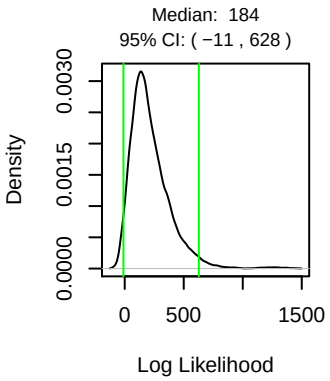

Mean: 49.7
St. dev.: 10.5
Frequency
0
1000
3000
20
40
60
80
Transistions from E to I

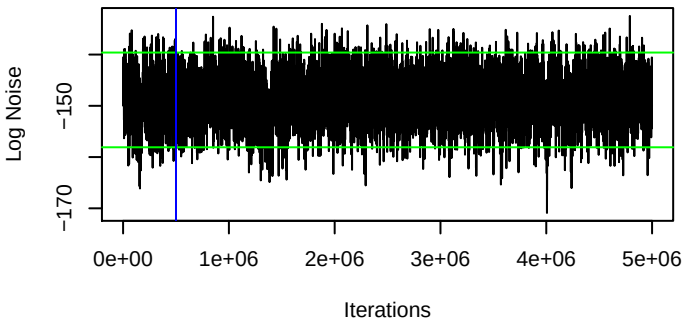

p_d
0.55
0.65
0.75
0.85
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Iterations
Median: 0.709
95% CI: (0.617 , 0.793)
Density
0
2
4
6
8
0.5
0.6
0.7
0.8
p_d

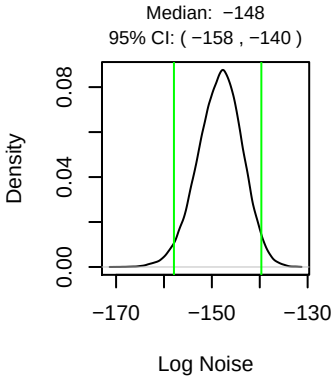

Sensitivity ELISA
0.65
0.75
0.85
0.95
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Iterations
Median: 0.814
95% CI: (0.72 , 0.892)
Density
0
2
4
6
8
0.6
0.7
0.8
0.9
Sensitivity ELISA


Sensitivity Culture
0.60
0.75
0.90
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Iterations
Median: 0.811
95% CI: (0.712 , 0.892)
Density
0
2
4
6
8
0.60
0.75
0.90
Sensitivity Culture


Log Likelihood
0
500
1000
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Iterations


Median: 184
95% CI: (−11 , 628)
Density
0.0000
0.0015
0.0030
0
500
1500
Log Likelihood


Log Noise
−150
−170
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Iterations


Median: −148
95% CI: (−158 , −140)
Density
0.00
0.04
0.08
−170
−150
−130
Log Noise

Appendix D

Programming approach

A large proportion of the work presented in this by computer code written by the author in C++ using an object oriented approach. The code used in Chapter 2 was inherited from Ross Davidson (SRUC) and extended to include a spatial component and some modifications to the disease transmission processes. The original model was published in Davidson et al. (2008). Chapters 3 and 4 use the same code which was written from scratch following a similar approach as the differential fecundity code used for Chapter 2. All code runs in a Linux environment and produces ascii output. The graphic throughout this thesis were produced using scripts in R (R Development Core Team, 2006).

A brief description of the structure of the most important algorithms used in this PhD are given below.

D.1 Differential fecundity algorithm

A brief outline of the differential fecundity algorithm is given below:

```
1. For number of runs
    1.1. Initialise population size and structure
    1.2. While time < equil_population_time
        1.2.1. Gillespie algorithm
    1.3. Introduce infectious individuals
    1.4. While time < equil_disease_time
        1.4.1. Gillespie algorithm
    1.5. While time < end_time
        1.5.1. Gillespie algorithm
2. Generate output files
```

The Gillespie algorithm consists of three steps:

```
1. Calculate event rates and total rate
2. Calculate time to next event based on total rate
3. Choose event type based on event rates
```

D.1.1 Performance

The performance of the differential fecundity algorithm depends primarily on the size of the population which this determines the total event rate. We have run the simulations in Chapter 2 on the EDDIE cluster at Edinburgh University. This has the advantage that our simulations are run as an exclusive process on a particular node and many simulations (for different parameter combinations) can be run simultaneously. The simulation for a single parameter combination for largest models (group carrying capacity $K = 100$, number of groups $N_g = 9$) take a few minutes to run on a single node on the eddie cluster. Generating the figures in this chapter involves simulating many thousands of parameter combinations and can take up to 24 hours to complete.

D.1.2 Example of parameter file for differential fecundity algorithm

```
##############################################################################

# Parameters for differential fecundity algorithm

# Lines containing parameters start with a description, followed by the
# parameter name in square brackets and the parameter value.
# parameter values: single parameter values or range (first;last;increment)
# Any lines starting with # will be ignored

##############################################################################

# Initialisation ------------------------------------------------------------------------
Number of social groups in one direction            [num_groups]             3
Number of social levels in each group (N_l)         [num_levels]             10
Average carrying capacity in a level                [av_carrying_cap]        10
Initial population in all groups (sum(n_i))         [initial_population]     350
Length of time for population level to relax        [equil_pop]              50.0
Initial number infected in all groups               [ninfection_seeds]       200
Time to relax the system post disease introduction  [equil_disease]          150.0
Time to then run after innoculation for averaging   [avg_time]               400.0
#-------------------------------------------------------------------------------------------

# Demographic parameters----------------------------------------------------------------
Birth rate coefficient (r_b)                        [birth_rate_coeff]       1.0
Death rate of infective individuals (mu_I)          [infective_death_rate]   0.1
Death rate of susceptible individuals (mu_S)        [susceptible_death_rate] 0.1
Rate of attempts to move (nu_m)                     [movement_rate]          0.0;1.51;0.02
Const for promotion acceptance prob (phi_p)         [promotion_const]        0.0;0.351;0.005
Prob failed promo leads to death (mu_p)             [challenge_mort]         1.0
Const for dispersal acceptance prob (phi_d)         [dispersal_const]        500.0
Prob failed dispersal leads to death (mu_d)         [disperse_mort]          0.0
#-------------------------------------------------------------------------------------------

# Hierarchy options ---------------------------------------------------------------------
# Hierarchy type: 0 = flat, 1 = linear, 2 = Gaussian, 3 = stepped
Type of hierachy                                    [hierarchy_type]         1
# Linear hierarchy parameters
y-intercept of hierarchy line (1=flat,0=sloped)(c)  [intercept]              0.0;1.0;1.0
#-------------------------------------------------------------------------------------------

# Spatial parameters--------------------------------------------------------------------
Proportion of movement attempts local promotion     [p_prom]                 0.0;1.0;1.0
#-------------------------------------------------------------------------------------------

# Disease transmission probabilities----------------------------------------------------
probability of vertival transmission (p_v)          [vtrans_prob]            0.25
Overall htrans strength Gamma                       [htrans_coeff_gamma]     0.05
```

```
Proport allocated to global transm (alpha)          [htrans_coeff_alpha]     0.0
Probability of infect on promotion attempt          [challenge_infect]       0.0
#-------------------------------------------------------------------------------------------

# Output control-----------------------------------------------------------------------------
number of runs over which to averages               [number_of_runs]         10
#-------------------------------------------------------------------------------------------
```

D.2 Inference algorithm

A brief outline of the inference algorithm is given below:

```
1. Read data
2. Initialise event sequence
3. Initalise parameters and traces
4. Calculate likelihood and sums
5. While n <= number of iterations
    5.1 While m <= number of changes to event sequence
        5.1.1. Propose a change
        5.1.2. Recalculate likelihood
        5.1.2. Accept or reject proposed event sequence
    5.2. Gibbs sampler for process parameters
    5.3. Metropolis-Hastings sampler for K and observation parameters
        5.3.1. Propose new value
        5.3.2. Recalculate likelihood
        5.3.3. Accept or reject proposed value
    5.4. Store parameters at end of trace
6. Write output files (traces and final states)
```

D.2.1 Performance

The performance of the inference algorithm depends on a number of factors, including the number of parameters being estimated, the number of changes made to the event sequence during each iteratiion, the amount of data included (i.e. length of observation period and number of groups) and the number of events within the observation period (note that larger populations tend to have higher event rates). The inference runs in Chapters 3 and 4 were done on the EDDIE cluster at Edinburgh University. In the most intesive runs, estimating all parameters on 15 groups, it can take up to 12 hours to 12 hours to complete 1 million iterations. Note that the parameters for the inference code (Section D.2.2) include a number of "tuning" parameters which affect the rate of convergence of the code. These include the number of changes to the event sequence in each iteration, the standard deviations for the proposal distributions for those parameters estimated using Metropolis Hastings samplers, the probability of making a particular type of change to the event sequence (add, remove, change timing, change initial state) and the probabilities with which events of each type are added.

D.2.2 Example of parameter file for inference algorithm

```
###############################################################################

# Parameters for MCMC inference

# Lines containing parameters start with a description, followed by the
# parameter name in square brackets and the parameter value.
# Any lines starting with # will be ignored

###############################################################################

# Run time parameters
#------------------------------------------------------------------------------
number of iterations                                    [nIts]          500000
number of proposed state space changes per iteration    [mIts]          100
model type (0=demography only, 1=including disease)     [ModelType]     1
thinning of output (every nth sample is stored)         [ThinOut]       50
initialise from end of existing run (0=new, 1=append)   [AppendRun]     1

# Switches for indicating fixed parameters (0=not fixed, 1=fixed)
#------------------------------------------------------------------------------
fix initial R                                           [FixR]          1
fix birth rate                                          [Fix_rb]        1
fix death rate                                          [Fix_mu]        1
fix mean of K                                           [Fix_MeanK]     1
fix sigma of K                                          [Fix_SigmaK]    1
fix detection probability                               [Fix_pd]        1
fix ELISA sensitivity                                   [Fix_SE]        1
fix culture sensitivity                                 [Fix_SC]        1
fix infection rate                                      [Fix_beta]      0
fix background infection rate                           [Fix_phi]       0
fix rate of transition EtoI                             [Fix_tauEI]     0
fix rate of transition ItoR                             [Fix_tauIR]     1

# Initial values
#------------------------------------------------------------------------------
initial birth rate                                      [Init_rb]       0.652
initial death rate                                      [Init_mu]       0.45
initial mean of carrying capacity                       [InitMean_K]    56.9
initial stdev of carrying capacity                      [InitSigma_K]   8.665
initial infection rate                                  [Init_beta]     0.02
initial background infection rate                       [Init_phi]      0.01
initial rate of transition EtoI                         [Init_tauEI]    0.5
initial rate of transition ItoR                         [Init_tauIR]    0.0
initial detection probability                           [Init_pd]       0.8
initial sensitivity ELISA                               [InitSensE]     0.8
initial sensitivity culture                             [InitSensC]     0.8
initial sensitivity R                                   [InitSensR]     0.1

# Prior Distributions
#------------------------------------------------------------------------------
# Note that we read the mean and variance for gamma  & beta priors and convert
# them to alpha and beta when reading them. For the beta distributions there is
# the option to choose U[0,1] and overide the parameter given.
#------------------------------------------------------------------------------
mean of gamma prior of rb                               [Mean_rb]       0.6
variance of gamma prior of rb                           [Var_rb]        0.01

mean of gamma prior of mu                               [Mean_mu]       0.3
variance of gamma prior of mu                           [Var_mu]        0.01

mean of normal prior of mean of K                       [Mean_Mean_K]   45
stdev of normal prior of mean of K                      [Sigma_Mean_K]  10
mean of normal prior of sigma of K                      [Mean_Sigma_K]  10
stdev of normal prior of sigma of K                     [Sigma_Sigma_K] 3

mean of gamma prior of beta                             [Mean_beta]     0.08
variance of gamma prior of beta                         [Var_beta]      0.0025
```

```
mean of gamma prior of phi                                [Mean_phi]      0.05
variance of gamma prior of phi                            [Var_phi]       0.0025

mean of gamma prior of tauEI                              [Mean_tauEI]    0.6
variance of gamma prior of tauEI                          [Var_tauEI]     0.0025

mean of gamma prior of tauIR                              [Mean_tauIR]    0.2
variance of gamma prior of tauIR                          [Var_tauIR]     0.0005

set prior for pd to uniform (1 = unif, 0 = use params)  [Unif_pd]       0
lower bound of uniform prior of pd                        [Min_pd]        0.61
upper bound of uniform prior of pd                        [Max_pd]        0.99
mean of beta prior of pd                                  [Mean_pd]       0.8
variance of beta prior of pd                              [Var_pd]        0.0025

set prior for SE to uniform (1 = unif, 0 = use params)  [Unif_SE]       0
lower bound of uniform prior of SE                        [Min_SE]        0.01
upper bound of uniform prior of SE                        [Max_SE]        0.99
mean of beta prior of ELISA sensitivity                   [Mean_SE]       0.8
variance of beta prior of ELISA sensitivity               [Var_SE]        0.0025

set prior for SC to uniform (1 = unif, 0 = use params)  [Unif_SC]       0
lower bound of uniform prior of SC                        [Min_SC]        0.01
upper bound of uniform prior of SC                        [Max_SC]        0.99
mean of beta prior of Culture sensitivity                 [Mean_SC]       0.8
variance of beta prior of Culture sensitivity             [Var_SC]        0.0025

# Proposal distributions for Metropolis Hastings samplers
#-------------------------------------------------------------------------------
standard deviation for proposal of mean of K              [PropSig_MeanK] 2.0
standard deviation for proposal of sigma of K             [PropSig_SigK]  1.0
standard deviation for proposal of pd                     [PropSigma_pd]  0.05
standard deviation for proposal of SensE                  [PropSigma_SE]  0.02
standard deviation for proposal of SensI                  [PropSigma_SC]  0.02

# Paramaters for generating alternative realisations of the state space
#-------------------------------------------------------------------------------
type of modification to make                              [ModType]       0
#switch (0=single events, 1 = pairs, matched)
probability of changing the initial state                 [ModInitProb]   0.01
probability of adding new event                           [AddProb]       0.05
probability of deleting an event                          [DelProb]       0.05
# Note: ShiftProb is calculated as 1-ModInitProb-AddProb-DelProb

probability of adding a birth                             [BirthProb]     0.4
probability of adding a death                             [DeathProb]     0.4
probability of adding an infection                        [InfectProb]    0.4
probability of adding a background infection              [BGInfectProb]  0.4
probability of adding a transition from E to I            [EtoIProb]      0.4
probability of adding a transition from I to R            [ItoRProb]      0.000001
# Note: need to make sure that these are possitive and add to 1, if sum neq
# 1 the probabilities get rescaled.....
```